MEASURING HUMAN SALIVARY AMYLASE COPY NUMBER VARIATION

Sugandha Dhar

Thesis submitted to the University of Nottingham for the degree of Masters by Research

September 2010

ABSTRACT

Copy number variations represent large scale genomic alterations varying from 1kb to 3Mb and are proposed as a driving force for genome evolution and variation. One such locus exhibiting copy number variation and genome evolution is salivary amylase, which is responsible for the digestion of starch in the human parotid glands. It was reported that since human salivary amylase gene (AMY1) copy numbers are correlated positively with protein levels, and also due to the correlation of high gene copy numbers with a starch rich diet, AMY1 had undergone adaptive evolution. Moreover there was recent evidence for gain of copy number in the human lineage concordant with the dietary shift which occurred during Neolithic times. The main focus of this study was the development of multiplex Paralogue Ratio Test (PRT) systems for measuring human amylase copy number. Despite having considerable identity between salivary amylase and pancreatic amylase copies, PRT was designed due to the exclusive association of a retroviral element with the salivary amylase gene copies, such that only salivary amylase copies were quantified. As retroviral elements are scattered throughout the genome therefore eliciting amplification from just test and reference without any contribution from alternative loci was a technical challenge in this project. However, creation of enough mismatches within the primer binding sites ensured that only the test and reference sequences were amplified. Discrimination by means of increasing annealing temperature also aided in exclusive amplification of test and reference sequences. Consequently, two PRTs and 1 microsatellite assay were successfully designed to measure copy number of AMY1 in Japanese and UK DNA samples. 14 copies was the highest copy number seen in the samples from both PRT systems. Reference samples specifying individual classes of copy number exhibited strong relationship with the copy numbers as determined by previous work. Repeat testing and clustering of copy number data points by both the individual systems of measurement ascertained the accuracy and reproducibility of the developed assays.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor Professor John Armour for his invaluable support and guidance. I would like to extend my special vote of thanks to Dr Jess Tyson for her encouragement and help. My sincere thanks also goes out to Tamsin Majerus, Dannie Carpenter, Suhaili Abu Bakar, Raquel palla, Somwang Jayakhantikul, Fayeza Khan, Ioannis Ladas. I would also like to thank my parents especially my father who has always helped me chase my dreams and stood by my decisions.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AMY1	Amylase (salivary)
AMY2	Amylase (pancreatic)
a-CGH	Array-Based Comparative Genomic Hybridisation
BAC	Bacterial Artificial Chromosome
BIR	Break Induced Replication Pathway
BSA	Bovine serum albumin
CGH	Comparative genomic hybridization
CNPs	Copy Number Polymorphisms
CNV	Copy number variations
CNVs	Copy Number Variants
ddNTPs	dideoxyNucleotide triphosphate
dNTP	deoxyribonucleotide triphosphate
DSB	double stranded break
ERV	Endogenous retrovirus
ECACC	European Collection of Cell Cultures
FISH	Fluorescent in situ hybridization
FoSTeS	Fork Stalling and Template Switching
LCRs	Low copy repeats
LCVs	Large-scale Copy Number Variations
LD	Low DNTP
LTR	Long terminal repeat
LINE	Long interspersed nuclear elements
MAPH	Multiplex Amplifiable Probe Hybridization
MLPA	Multiplex Probe Ligation Assay
NAH R	Non-Allelic Homologous Recombination
NHEJ	Non Homologous End-joining
PCR	Polymerase Chain Reaction
PPPP	PRT primer picking programme
PRT	Paralogue Ratio Test
QF-PCR	Quantitative Fluorescent PCR
RT-qPCR	Real Time Quantitative PCR
SDs	Segmental Duplications
SDSA	Strand Dependent Synthesis Annealing
SINE	Short interspersed nuclear elements
SNP	Single Nucleotide Polymorphism

CHAPTER1: INTRODUCTION

1.1: The Human Genome

The human genome refers to the total DNA content or genetic information present in human cells. It can be divided into the Nuclear Genome and the Mitochondrial Genome.

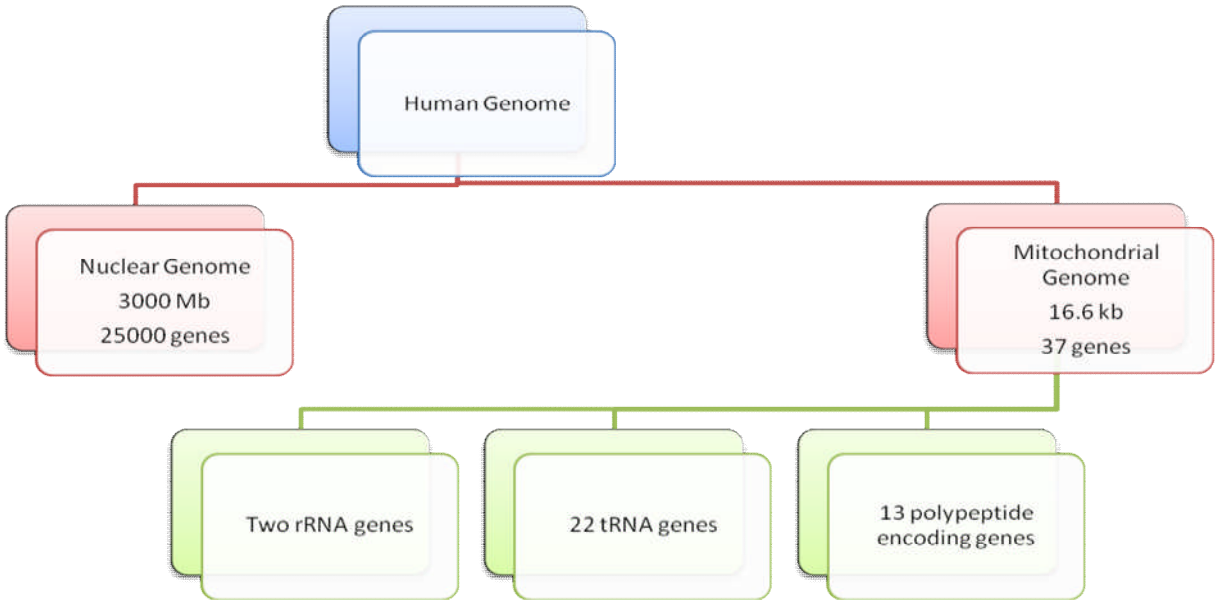


Figure 1. General overview of the human genome.

1.1.1: Mitochondrial Genome

The human mitochondrial genome is primarily composed of circular double-stranded DNA and sometimes triple-stranded structures are also seen (Clayton et al, 1992). Its nucleotide structure has been decoded (Anderson et al, 1981). The two strands designated as H (Heavy) and L (Light) differ in base compositions with H strand being rich in guanines and L in cytosines and the third strand contributing to the triple stranded DNA appearance represents the replication “D-loop” of 7S DNA which is a segment of H strand.

Briefly, mitochondrial genome is 93% coding, has very few repeats and is relatively simpler than nuclear DNA in terms of its genomic organisation which may also be indicative of its endosymbiont origin. It is characteristised by the absence of introns, genome length of 16569 bp and presence of only 37 genes which produce RNA and mitochondrial peptides; therefore it uses the nuclear genome for synthesising other requisite components and imports them.

Briefly, out of the 37 genes, 24 genes synthesise RNA products (tRNA, rRNA, 23S RNA and 16S RNA) and 13 genes are responsible for the production of mitochondrial ribosomal polypeptides. 22 tRNA molecules recognise a total of 60 codons and account for polypeptides synthesised within the mitochondria. Any defect within the replication, transcription, translation and repair system or associated dysfunctions may lead to mitochondrial diseases.

Interestingly, transmission of mitochondrial genes is only through the female and not through the male as seen in humans, paramecium and snails. However for some species it is also paternally driven especially in mussels (Meusel et al, 1993), honey bees (Fontaine et al, 2007) and fruit flies (Kondo et al, 1992).

1.1.2: Nuclear Genome

The nuclear genome includes the following elements which code for 25,000 genes.

1.1.2.1: Genes and Regulatory Elements

The nuclear haploid human genome comprises 3 billion base pairs of euchromatic DNA utilized for human genome project and 200 kb of constitutive heterochromatin which is transcriptionally inactive and unevenly distributed. Out of 3 billion bp, 4.5% show conservation including 1.5% protein coding sequences and 3% of regulatory and untranslated sequences. Most of the DNA is translated (90-95%) while the remaining 5-10% is untranslated and transcribed as RNA. Non-coding DNA consists of tandem repeats or dispersed repeats formed by segmental duplication or retrotransposition of RNA transcripts. It is also now established that DNA is divisible into different categories of repeated sequences through cot curve analysis which involves shearing, denaturing, reannealing DNA and then measuring the percentage of single-stranded molecules left at different time intervals.

The latest evaluation suggests that there are 23438 unique protein coding genes with 500000 exons; they give rise to 140000 different proteins with more than 6000 RNA coding genes, 6407 pseudogenes and 183 genes performing unknown function (Ensembl, September 2009).

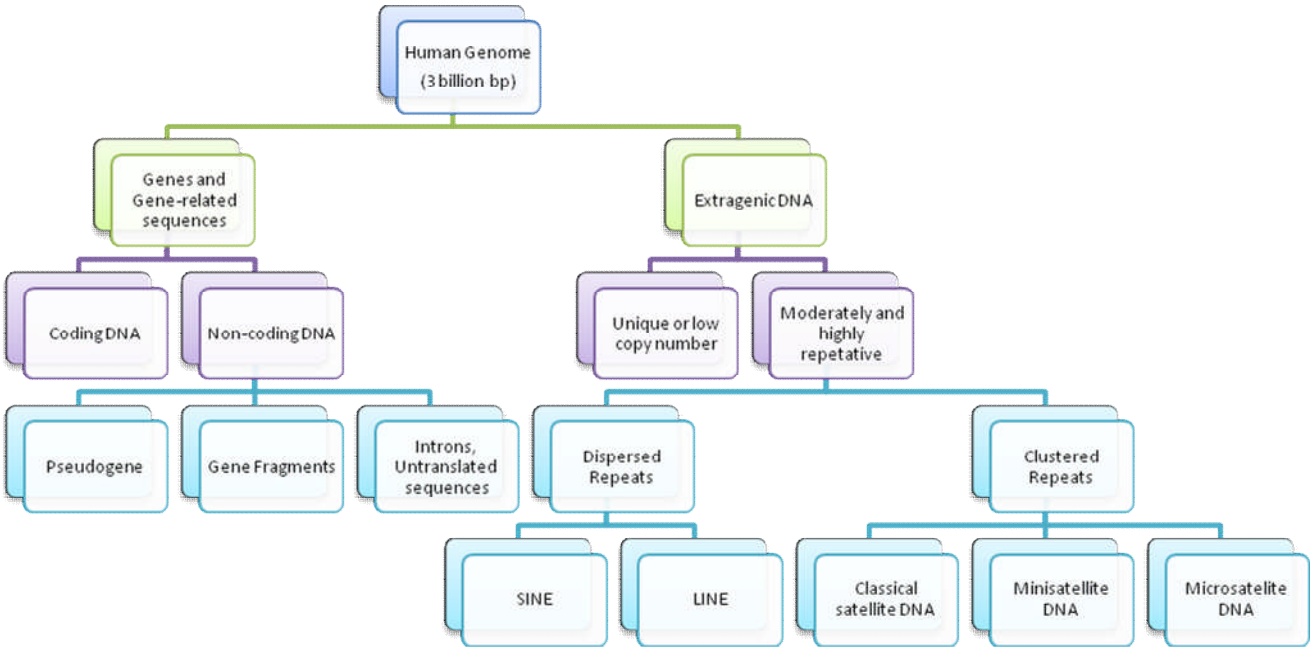


Figure 2. Different components of the human genome influencing interindividual variations (genetic differences and corresponding phenotypic manifestations). The three billion base pairs of DNA can grouped into genes, related sequences and extragenic DNA.

1.1.2.2: Pseudogenes

As the name suggests, pseudogenes represent defective copies of genes which which are unable to produce a fuctional protein either due to the presence of stop codons, frameshift mutations, absence of regulatory elements or presence of truncated sequences within the DNA sequence which changes its functionality. Pseudogenes may be non-processed and processed pseudogenes. Non processed pseudogenes or duplicated pseudogenes are generated by the duplication of the ancestral gene which maintains the original function but the duplicated copies acquire mutations and subsequently looses it original function. Processed pseudogenes or retrotansposed pseudogenes are produced by the retrotransposition of mRNA

transcripts back into the genomic DNA through reverse transcriptase. They lack introns, regulatory elements but possess poly A tails.

1.1.2.3: Repeat Elements

Repeat elements were first recognised by Britton and Kohne (1968) in their studies pertaining to reassociation rate of denatured fragments and it was gradually observed that they constitute more than half of the genomic material for humans (Lander et al, 2001). Now it is established that even though these elements do not seem to be involved in transcription and translation directly, they do influence transcript and the protein product synthesized and are mediators of evolutionary change, genomic length, position and mobility. Function and copy number are the parameters by which repeat elements are characterized.

1.1.2.4: Dispersed Repeat Elements

Dispersed repeat elements are dispersed all over the genome. They include short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE) and account for 45% of human genome (Lander et al, 2001). Transposable elements are characterised by their ability to move to different locations and their ability to propagate themselves. At first, transposons were also called selfish DNA as they replicate independently as compared to rest of the genome; however new studies suggest that natural selection is responsible for their propagation (Vinckenbosch et al,2005). Transposition is a mechanism by which these nuclear elements are able to insert themselves permanently within the genome. Transposition can occur through the formation of an RNA intermediate (RNA transposons which are restricted to a particular genomic location and its copies move) or DNA intermediate (DNA transposons which move by cutting the specific sequence and that sequence moves) and consequently give rise to different elements within the human genome.

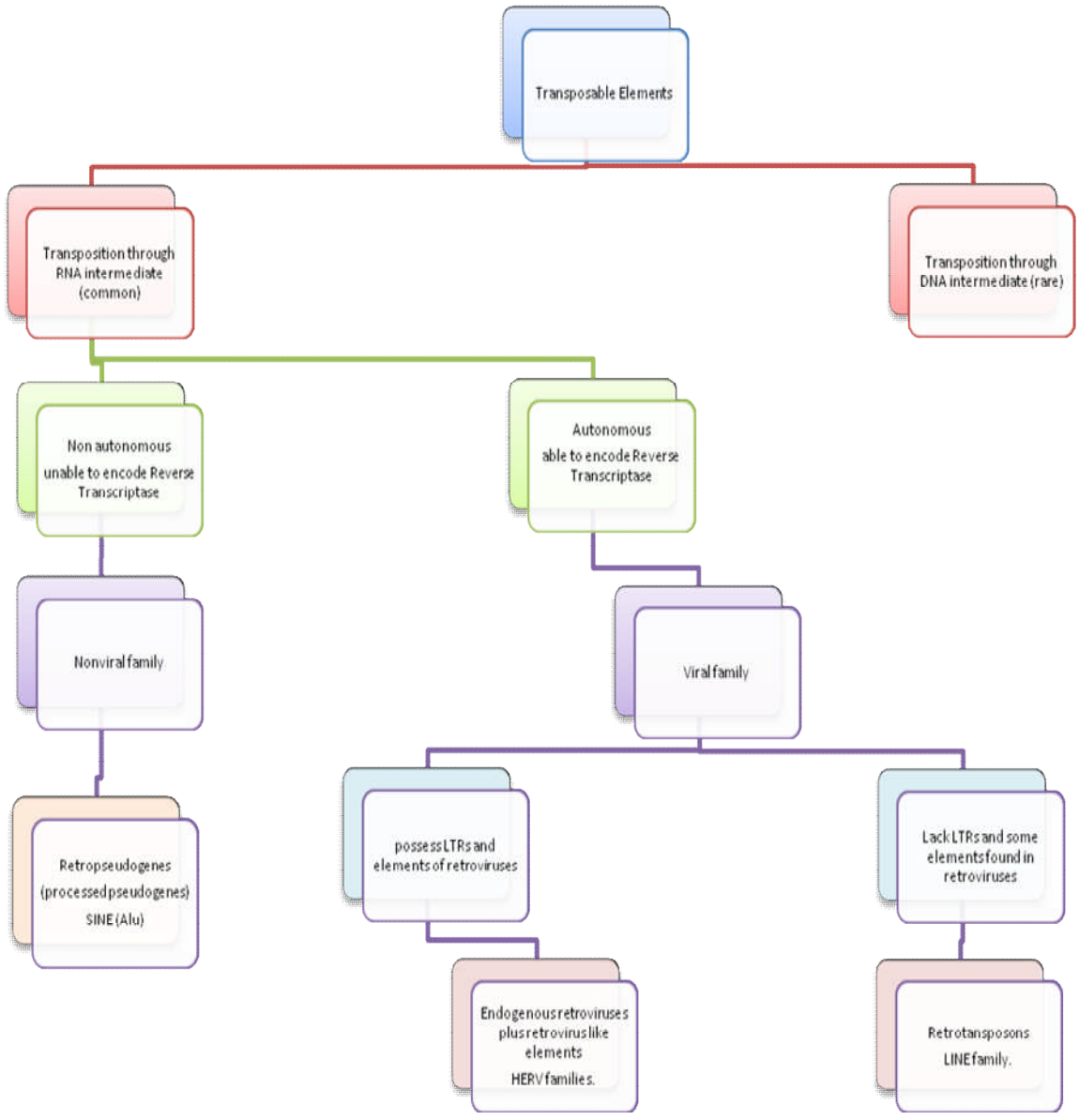


Figure 3. Classification of the different transposon families in humans based on the mode of integration into the target site.

1.1.2.5: Tandem repeat Elements

Tandem repeats constitute highly repetitive DNA sequences usually of 1 to 500 nucleotides in length. These represent short sequences in long tandem arrays often located near heterochromatin area and also appear as secondary constrictions in metaphase chromosomes. Broadly, tandem repeats are divisible into three types of repeated sequences depending on their density with respect to the caesium chloride density gradient, which is further dependent upon size of the repeated sequence.

Minisatellites (VNTRs) have repeat units 9-64 bp in length, are often hyper variable and are found near the telomeres.
Microsatellites are 2-6 bp repeats which are less than 200 bp in length.
Satellite DNA (Singer et al, 1982) is mostly found near the centromeric heterochromatin.

1.1.2.5.1: Segmental Duplications

These low copy repeats occur at different regions within the genome, share more than 90% sequence identity amongst each other (Eichler et al, 2001) and constitute 5% of the human genome. These DNA stretches range between 1-400 kb and are often associated with regions of chromosomal instability and hence act as mediators of chromosomal rearrangement mechanisms. Recent studies suggest that segmental duplications are 31% concentrated around pericentromeric regions, 2% at subtelomeric and 67% around interstitial regions and might also be restricted around one or different genomic locations (Koszul et al, 2009)

1.1.2.6: General Structural Variation

The human genome project has revealed that generally all humans are 99.9% identical and at any one time there can be up to 0.1% differences. Essentially this 0.1% of 3 billion haploid bps which influences genetic variation and disease susceptibility. This genetic variation may be polymorphic, that is found in more than 1% of the population or due to more recent mutation which reflects the unique differences in DNA sequence of an individual that are rare in a population.

Depending on the frequency and scale of its extent, this genetic variation is classified into the following categories:

Variation Type	Definition	Frequency in the human Genome	References
•SNP	•Single base pair variation found in >1% of chromosomes in a given population.	•~10 million SNPs	•International Human genome sequencing consortium,2004; Kruglyak et al,2001;Lander et al,2001.
•Insertion/ Deletion variant (Indel)	•Deletion or insertion of a segment of DNA. Includes small polymorphic changes and large chromosomal aberrations.(Indels). >1 kb in size called CNVs	•~1 million insertion/deletion polymorphisms>1 bp	•Weber et al, 2002;Dawson et al,2001
•Microsatellites	•Sequences containing variable numbers of 1-6 bp repeats totalling <200 bp in length.	•1 million microsatellites in the human genome, accounting for ~3% of the sequence	•Lander et al,2001, Liee et al, 1989; Ellegren et al, 2004.
•Minisatellite and VNTRs	•Polymorphic sequence containing 20-50 copies of 6-100 bp repeats.	•~150000 minisatellites, of which ~20% are polymorphic	•Nakamura et al,1987; Jeffreys et al, 1985; Naslund et al, 2005.
•Multisite Variant	•Single nucleotide variant with complex characteristics due to CNV or gene conversion.	•MSVs number currently unknown	•Fredman et al,2004.
•Intermediate-sized structural variant	•Gain or loss of a DNA sequence>8 kb in size also includes inversion breakpoints.	•297 ISVs	•Tuzun et al, 2005.
•CNV:CNP	•Copy number change > 1 kb. If the frequency is >1% then called CNP. LCVs are CNVs ~50 kb in size or greater	•Frequency of CNVs unknown. Larger CNVs(>50 kb)	•Iafrate et al, 2004; Sebat et al, 2004; Feuk et al, 2006.
•Inversion	•Rearrangement due to which DNA fragment is reversed in orientation.	•microscopically setectable inversion frequencies 0.12-0.7% (pericentric) and 0.1-0.5% (paracentric); submicroscopic unknown	•Van et al, 1983; Gardner et al, 2004.
•Translocation	•Rearrangement which causes DNA fragment to be attached at a different chrmosome.	•1/500 heterozygous for reciprocal translocation and 1/1000 for robertsonian translocations	•Van et al, 1983; Gardner et al, 2004; Warburton et al, 1991.
•Unbalanced rearrangements	•Rearrangement which leads to a net gain or loss of DNA, unbalanced in nature	•unbalanced rearrangements occur in ~1/15000 live births.	•Nassbaum et al,2004.

Table 1. Classification of different types of variation present within the human genome based on the frequency and scale of its extent.

1.1.2.6.1: Submicroscopic Structural Variation

Structural variation describes genomic alterations involving insertions, deletions, duplications (having a quantitative effect) and inversions, translocations (having positional effect) and are larger than 1000 bp. The existence of submicroscopic structural variations (~1 Kb to ~3 Mb) was one of the major revelations of the human genome project (Lander et al, 2001; Sebat et al, 2004; Redon et al, 2006) and 30% of the human genome was subject to structural variation (Zhang et al, 2009).

1.1.2.6.2: Copy Number Variation

Copy number variations correspond to the changes (>1 kb) in number of copies of DNA sequences in comparison to a reference chromosome (Feuk et al, 2006). Some of these variations are mentioned in the database of genomic variants having 89,427 entries covering 12% of the genome (Redon et al, 2006) and as compared to SNPs, a higher mutation rate (100-1000 times) for CNVs has been predicted (Van et al, 2005). Moreover, CNVs also account for 0.37% of the differences between different human genomes which is reflective of their importance in evolution and genetic diversity (Levy et al, 2007; Kim et al, 2009). However, complications arise while measuring and assigning definite integer copy number within a copy number variable region due to recurring somatic mutations occurring throughout lifespan of an individual and the existence of different copy numbers in different tissues. CNVs can also be used for identification as demonstrated by the existence of different CNVs prevalent in monozygotic twins (Piotrowski et al 2008, Bruder et al, 2008).

1.1.2.6.2.1: Evolutionary potential of copy number variations

Copy number variations are not the exclusive feature of *Homo sapiens* but have also been detected in mice and flies. In some arthropods like Drosophila, most CNVs seem to be selected against as they are apparently deleterious (Dopman et al, 2007). Selection has been a major force in this case. Conversely, in humans Rhesus negative allele in mothers which causes erythroblastosis fetalis primarily due to Rhesus positive allele of fetus appears not to be removed by selection.

Copy number variations do possess the capability of influencing the translated product through mRNA and thus have an effect over fitness of an individual. However the correlation between copy number and expression is not always linear as there are other factors which also tend to influence expression like the presence of sequence variants within individual copies or interaction between copies, accessibility of DNA particularly with regards to regulatory elements or methylation etc.

For example, it has now been established that each copy of the salivary amylase genes is associated with an upstream promoter which is responsible for its expression which predicts that there is a correlation between gene copy number and expression levels (Perry et al,

2007). However at the same time, this does not mean that changes in gene copies would always correlate with gene expression. For example; OPN1LW and copy number variable OPN1MW are the two genes which code for red and green opsin visual pigment on chromosome X (Clayton et al, 1992). Mutations in OPN1MW may lead to colour blindness. However colour blindness arises when the gene located nearest to the promoter region is disrupted irrespective of the total number of copies of genes (Winderickx et al, 1992). Similarly, expression in α globin genes is not proportional to copy number due to competition between NME4 gene and globin genes for its enhancer sites such that any deletion of α globin genes results in increased expression NME4 gene (Higgs et al, 1989;Lower et al, 2009).

1.2: Mechanisms behind Structural DNA Exchanges

Changes in structure of chromosomes are mediated by gain or loss of copy number which eventually ends up joining two formerly separated DNA sequences and leads to DNA rearrangements. Genomic rearrangements involving gain, loss or disruption of dosage sensitive gene influence the phenotype of an organism through duplication, deletion, position effects and gene fusion or conversion events and may lead to genomic disorders (Lupski et al, 2005; 2006). Since DNA rearrangements define DNA changes involving thousands to millions of base pairs, their formation mechanisms were speculated to be different from monogenic point mutations which are usually formed as a consequence of DNA replication errors and repair, but recent evidence suggests that some disorders may be caused due to errors of replication and repair particularly involving homologous DNA strands (Lupski et al, 2005; Stankiewicz et al, 2002). Genomic instability is created due to the occurrence of repeat elements clustered together at a specific location which often incites chromosomal rearrangements to occur. Accordingly, human genome rearrangements may be divided into two major groups:

1. *Recurring Rearrangements*

These refer to rearrangements that have fixed break points, that is they are within the same interval in different individuals.

2. *Non Recurring Rearrangements*

These refer to rearrangements having distinct breakpoints but share a common region of overlap called the smallest region of overlap (SRO).

Genomic rearrangements in the human genome occur through these four mechanisms:

1. Homologous Recombination
2. Non Homologous End- Joining
3. Fork Stalling and Template Switching
4. L1 Retrotransposition

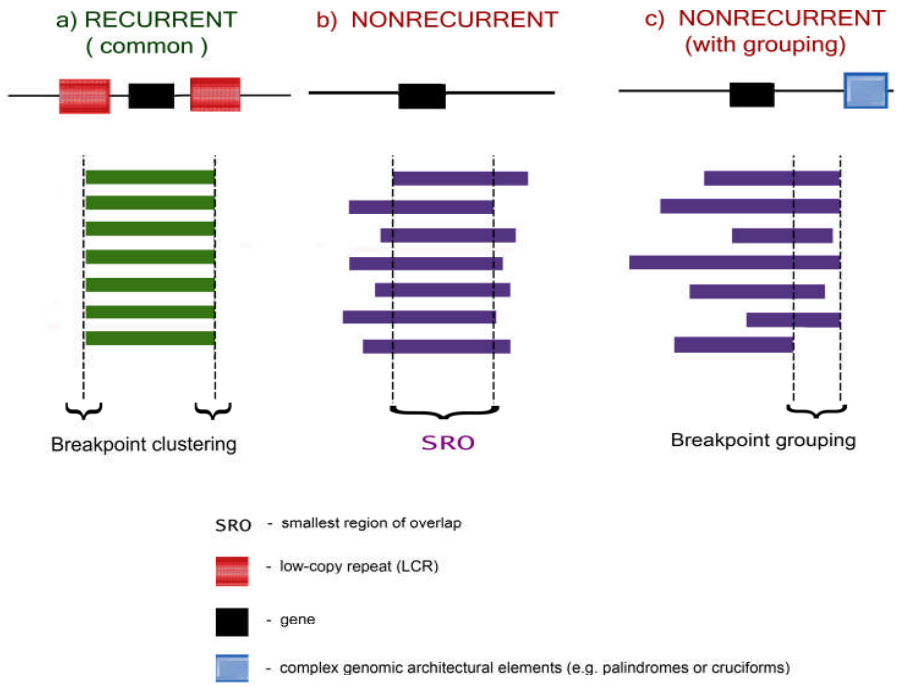


Figure 4. Characteristics of recurrent and non-recurrent genomic rearrangements with emphasis on smallest region of overlap.

1.2.1: Homologous Recombination

LCRs correspond to region specific DNA segments usually of 10 to 300 kb in size and exhibiting > 95% to 97% similarity amongst themselves (Bailey et al, 2006). Instead of the usual allelic copies, non-allelic copies exhibiting high sequence similarity are misaligned together in mitosis or meiosis and a subsequent crossing-over between them could result in genomic rearrangements. Therefore, non-allelic copies are often called substrates of Homologous Recombination. Non Allelic Homologous Recombination (NAHR) mediated by

segmental duplications, accounts for most large recurrent rearrangements. Non-recurrent events are activated by segmental duplications. However, some rare non-recurrent rearrangements are caused by some highly homologous repetitive sequences (Alu, LINE). Since the breakpoints of rearrangements have been mostly found to be associated with instability in the genome at specific regions having LCRs, it is known accepted that rearrangements are not a random phenomenon. Homologous recombination is a characteristic feature of meiosis in diploid organisms whereby crossing over enhances the genetic diversity through exchange of genetic material and is critical for the correct disjunction of chromosomes during metaphase. During the S phase of mitosis, homologous recombination between identical sister chromatids may be used to repair double strand breaks, therefore rearrangement caused by NAHR is not an exclusive feature of germ cells (Darai et al, 2008; Fridlyand et al, 2006). Most often NAHR is responsible for recurrent genomic rearrangements and occurs between two low copy repeats or segmental duplications (Shaw et al, 2004). Models of Homologous Recombination include:

1. ***Double Holliday Junction Pathway***
2. ***Strand Dependent Synthesis Annealing***
3. ***One Ended DSB Repair***

1.2.1.1: Double Holliday Junction Pathway and Strand Dependent Synthesis Annealing

Two-ended Strand-Break Repair can result either in Double Holliday Junction or in Strand Dependent Synthesis Annealing. First, double-stranded breaks are created in the DNA due to exposure of a mutagen like UV, the 5’ ends of DSB are then removed from 3’ overhanging tails. 5’ ends gets resected because they cannot retemplate back as replication only occurs from 3’ to 5’. Coating with Rad51 in eukaryotes then catalyses the invasion of homologous sequence by one or both 3’ ends ultimately generating a D Loop. Next step involves priming of the D loop with the chromatid and DNA synthesis occurs. Depending on the resolution through endonucleases and ligases, this would eventually lead to a non cross- over or a cross over being produced. Resolution can be vertical or horizontal but for producing recombinants for Two-ended DSB Repair, one end has to be resolved vertically and the other horizontally so that a switch over can occur. If both cuts are resolved horizontally (or vertically) then non-recombinants are produced (Figure 5).

1.2.1.2: Synthesis Dependent Strand Annealing (SDSA)

SDSA is a slight modification of the Double Holliday Junction pathway and is initiated in the same double Holliday junction pathway manner, the difference being in the polymerase extending step. In this case, helicase first separates the double helical DNA formed by the invading and synthesized strand. The invading strand then encounters the second end from DSB and anneals through complementary base pairing. The second end is extended by DNA synthesis and is ligated later after completion (Lupski et al, 2006) (Figure 5).

1.2.1.3: One ended double stranded break repair

Includes one stranded break and involves Break induced replication pathway. Break induced replication pathway (BIR) involves a replication specific enzyme called DNA helicase. DNA helicase is an enzyme responsible for converting dsDNA into ssDNA, but when it encounters a nick in the template strand, collapsed replication forks begin the BIR pathway. This is a modification of SDSA, in the sense that both have invasion from 3' end but in this case particularly extension of both leading and lagging strand occurs. However unlike in SDSA, the separated 3'end is not able to bind to a complementary second end to anneal. Reinvading by this 3' end occurs and is extended by the low processivity of the replication fork. This trend is followed until a more processive replication fork is formed (Hastings et al, 2009). During crossing over and non crossing over as well, it is evident that there would be patches of gene conversion events observed. Unequal crossing over may result in deletion (DiGeorge Syndrome), duplication (dup22q11) or inversion (Haemophilia). Sometimes mutation sequence from a non functional pseudogene is inserted into the functional gene through gene conversion (Figure 8).

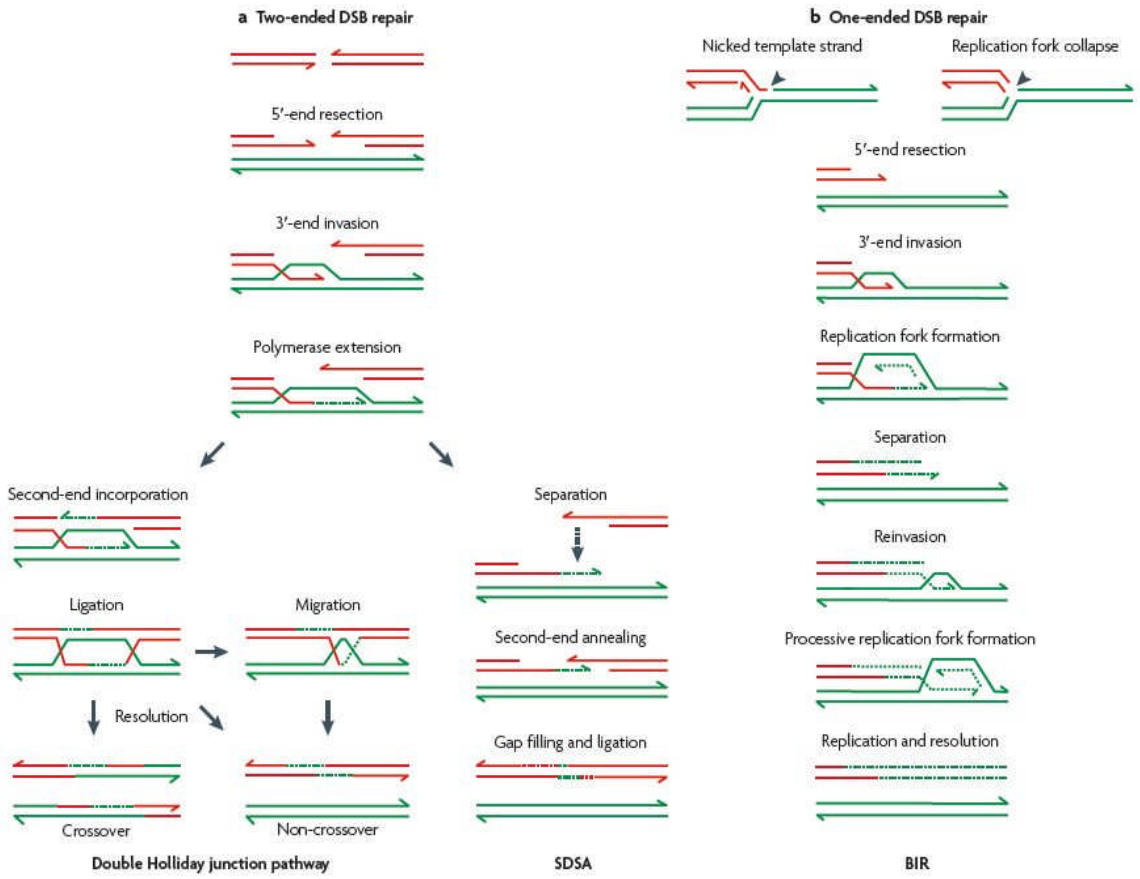


Figure 5. The potential mechanisms of Homologous recombination.
Source Mechanisms of change in gene copy number (Hastings et al, 2009).

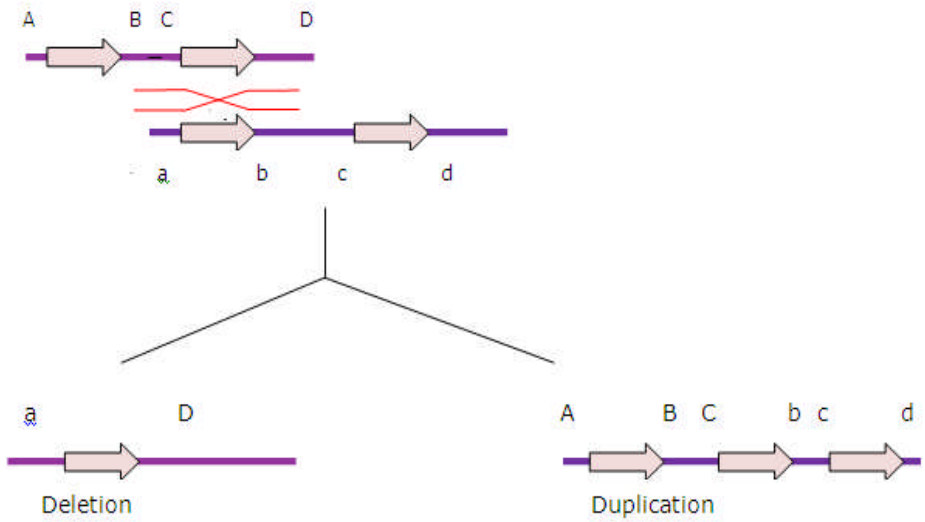


Figure 6. Orientation of different segmental duplications which eventually determine the fate and type of aberration.

Recombination between different direct repeats results in deletion and duplication highlighted by the above diagram.

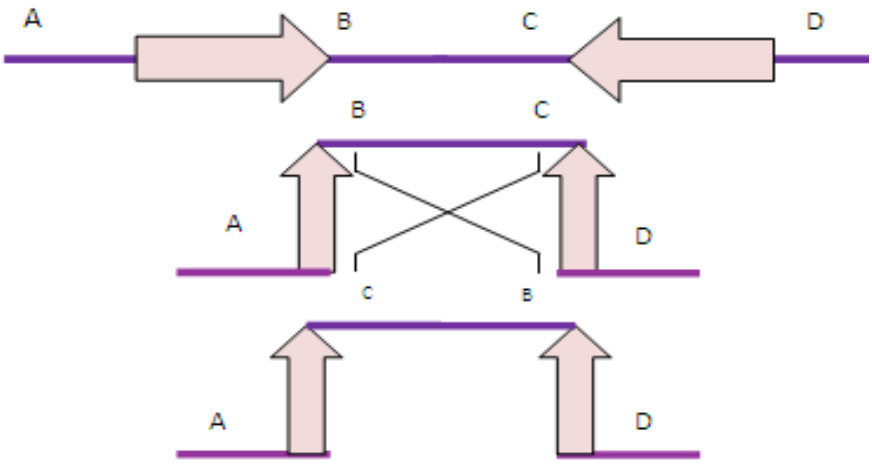


Figure 7. Inversion of different segments of DNA.
A, B, C and D represent different segments of DNA; out of which segments B and C get inverted due to recombination between inverted repeats highlighted by arrows.

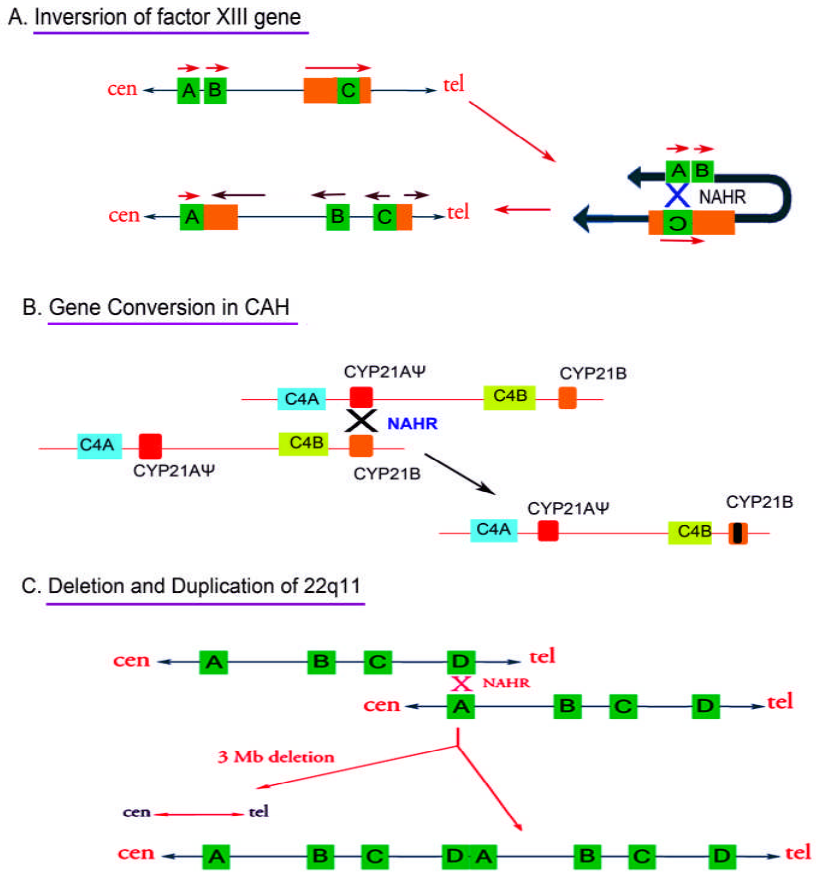


Figure 8. NAHR mechanism of structural DNA exchange causing disorders including deletion at 22q11, inversion leading to haemophilia A and Gene conversion in Congenital adrenal hyperplasia.

Although the role of NAHR as a mechanism behind structural DNA exchange has been uncovered, recent research focuses on differences in recombination frequency and homology length requirement between males and females and also between meiosis and mitosis (Lam et al, 2006; 2007).

1.2.2: Non Homologous End-joining (NHEJ)

LCR mediated NAHR does not explain all cases of genomic rearrangements. NHEJ molecular mechanism gives rise to non-recurrent chromosomal rearrangements with scattered break points (Figure 9). Non-homologous end joining is used to repair somatic double stranded DNA breaks primarily during G0, G1 and early S phase. It involves joining of the broken DNA strands without the use of a homologous template and utilizes very short homologous sequences (Microhomologies) to guide the repair. These micro homologies are often present as single-stranded overhangs in the ends of double stranded breaks. If they are not originally present in the breakpoint then overhangs are created by removal of a few bases, which allows complementary base pairing to occur. Hence, the hallmark of this mechanism is the deletion or insertion of a few bases around the breakpoint. Ku proteins are often utilised so as to bind to free DNA ends and promote the alignment of the two DNA ends along with recruiting enzymes. Kinases are often used for processing of ends and ligase for final ligation (Lieber et al, 2003, 2008).

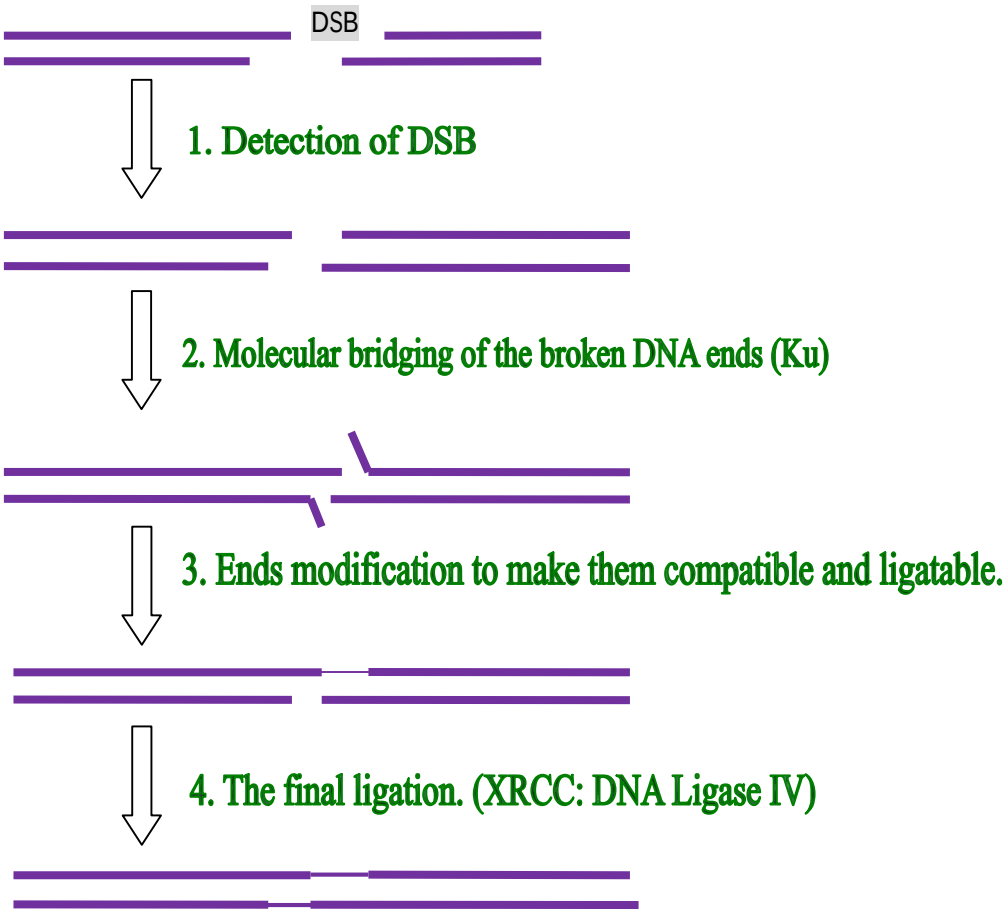


Figure 9. DSB produced due to disruption of the phosphodiester backbone of the double helix. After the detection of DSB, ku proteins align the DNA ends and protects them from degradation. This protein often engages DNA protein kinases to make the ends ligatable. artemis solely possesses 5' exonuclease activity. Artemis and DNA Pkcs possesses 5' and 3' overhang endonuclase cleavage activilty which means that this enzyme can trim 5' overhangs with a strong preference for the site that blunts the end and incontrast 3' overhangs are trimmed with a preference to leave a 4 or 5 nucleotide single stranded overhang.

1.2.3: Fork Stalling and Template Switching (FoSTeS)

This mechanism represents a replicative non-homologous repair system of DNA exchange. With the advent of more sophisticated technology, complex details of genomic rearrangements can be observed which has been the basis for propounding the FoSTeS model as a potential mechanism for some DNA exchanges. Rearrangements causing certain diseases like PMD (Pelizaeus-Merzbacher Disease) could not be fully explained on the basis of NAHR and NHEJ mechanism (Lee et al, 2007).

Accordingly, during replication when replication fork halts at a specific position, the template releases the lagging strand and anneals via 3' end homology to another replication fork which is in the nearest vicinity and restarts the DNA synthesis. Invasion and annealing are dependent upon the microhomology between invaded and original site. Deletion occurs if invasion of a new strand is forwardly directed and duplication is common in situations involving backward invasion. Thus the orientation of the replication fork would be dependent on the stand (leading or lagging) which underwent invasion and this invading step could be repeated multiple times which would ultimately reflect low processivity of polymerase (Lee et al, 2007).

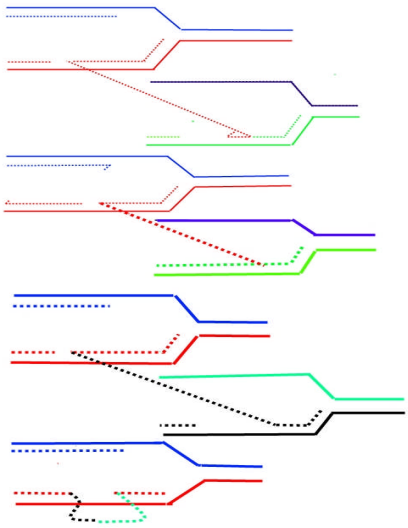


Figure 10. Fork stalling and template switching mechanism.
1. After the original stalling of the replication fork, the lagging strand disengages and anneals to a second fork via micro homology.
2. Extension of primed second fork and DNA synthesis.
3. The tethered original fork with its lagging strand may invade a third fork and this could occur several times before.
4. Resumption of replication on the original template.

1.2.4: L1 Retrotransposition

Retrotansposons represent the human genetic elements which insert their extra copies throughout the genome through copy and paste mechanism. LI Elements constitute almost ~20% of mammalian genomic DNA content. Most of these are retrotransposition incompetent because of truncated L1 copies but about 150 full length L1 elements are present within the human genome (Lander et al, 2001; Waterston et al, 2002; Goodier et al, 2001).

Full length non-LTR autonomous L1 is about ~6kb long and consists of a 5' UTR region containing an internal RNA polymerase II (RNAP II) promoter (Swergold et al, 1990) , two open reading frames (ORF 1 and ORF 2) and a 3' UTR containing a poly A tail signal. ORF 1 codes for an RNA binding protein and ORF 2 codes for protein with endonuclease and reverse transcriptase function (Babushok et al, 2007). This constitutes the equipment for Target primed Reverse Transcription and L1 elements are the sole autonomous transposon elements in the human genome.

Because of TPRT and decay over time, most of the L1 copies become disrupted by truncations, internal rearrangements and mutations (Lander et al, 2001). There is evidence of more than 500000 L1 copies in the human genome, out of which fewer than 100 are functional (Brouha et al, 2003).

1.2.4.1: The Retrotransposition Cycle

RNA polymerase II mediates the transcription of L1 locus through an internal promoter which in turn directs transcription initiation at the 5' end of L1 transposon (Lavie et al, 2004). Because of this internal promoter, RT is able to generate autonomous duplicate copies at different locations in the genome.

The transcript is then transferred to the cytoplasm where ORF 1 and ORF 2 are translated. Since both proteins produced show a *cis* preference (Wei et al, 2001) therefore they associate with the transcript that encoded them to produce a ribonucleoprotein (RNP) particle. This RNP is then transported back into the nucleus by a mechanism which is not understood properly until now.

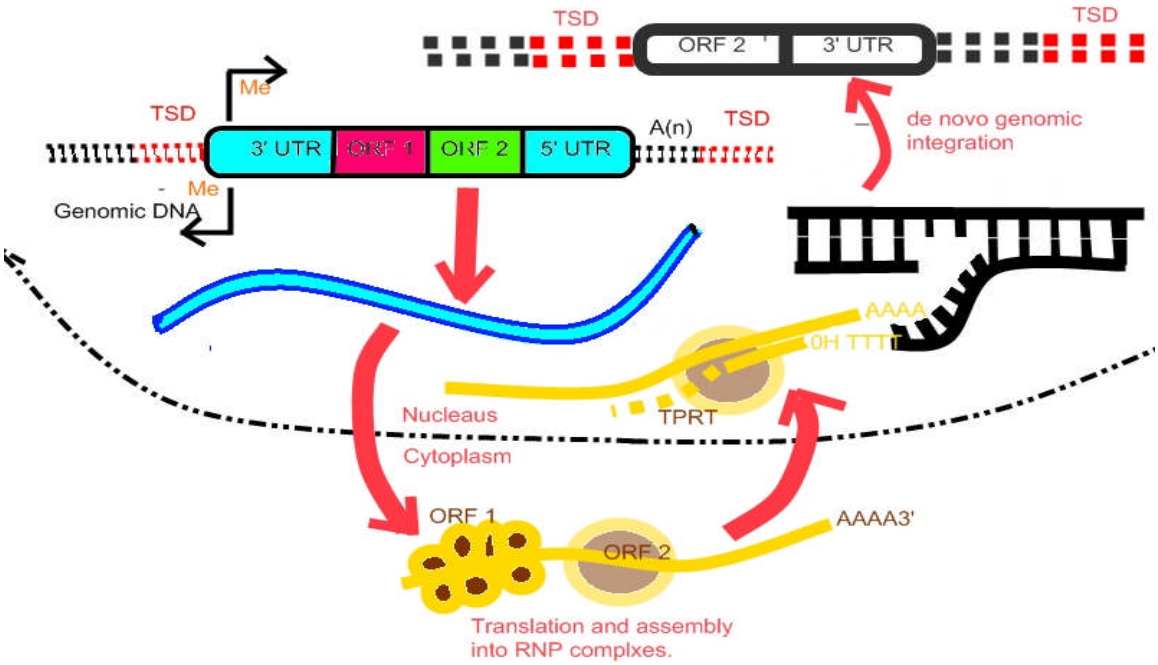


Figure 11. The retrotransposition cycle leading to de novo genomic integration.

The integration of L1 element into the genome is thought to occur through "Target – primed Reverse Transcription" (TPRT) (Feng et al 2002; Cost et al, 2002). During TPRT, L1 endonuclease cuts the first strand of target DNA, generally between T and A at 5'TTTTAA3' consensus sites (Jurka et al, 1997). Consequently, the free 3'OH liberated as an aftermath of the cut at the first strand, is used to prime reverse transcription of L1 RNA by L1 Reverse Transcriptase. After cutting the second strand of target DNA, 3' OH generated is used to prime second strand synthesis producing Target site duplications.

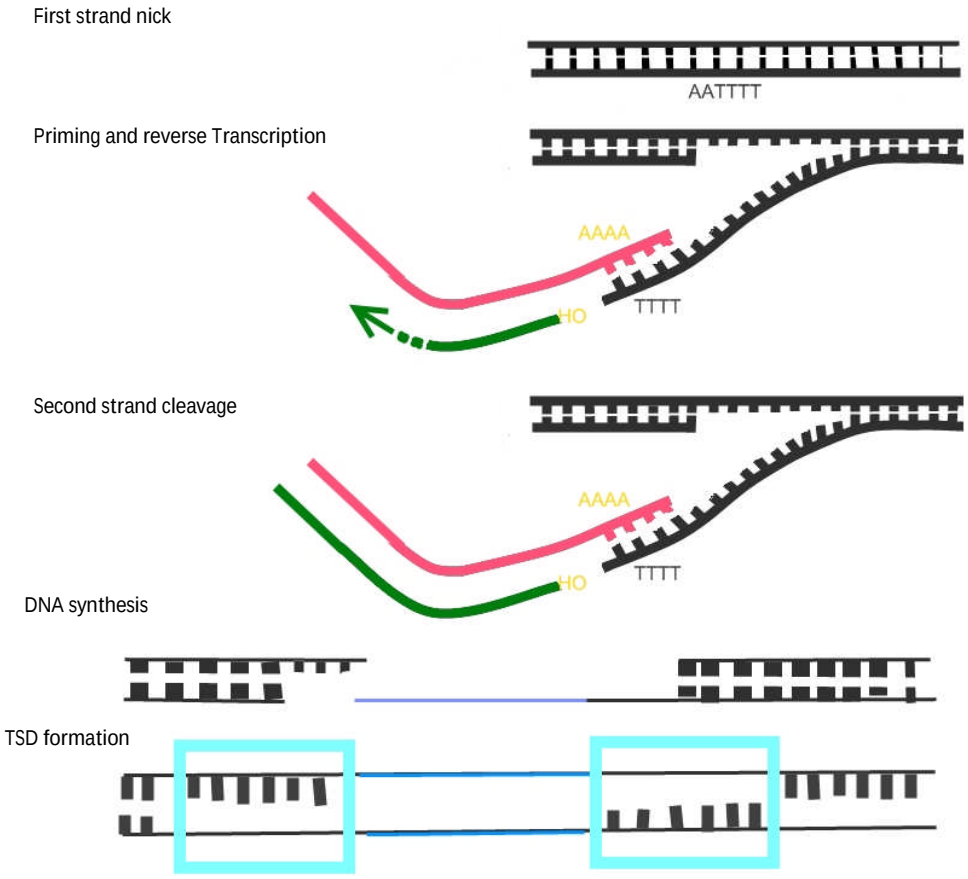


Figure 12. Target primed reverse transcription occurring in retrotransposition cycle. TPRT eventually leads to the propagation of retrotransposon transcripts throughout the human genome. LINE elements undergo transposition through the transcription of LINE elements into RNA which further codes for an RNA binding protein and a multifunctional protein with reverse transcriptase and endonuclease acidity. After the association of the above mentioned proteins with LINE RNA, the endonuclease protein nicks the DNA at a poly T rich site which eventually base pairs with poly A sequences in LINE RNA. LINE RNA is copied by the reverse Transcriptase into its DNA copy which is covalently attached to the target DNA. Template DNA copy then synthesises a second DNA strand and hence the target DNA produced at the flanking ends is filled by the generation of target site duplications (TSDs).

1.3: Methods for detection and measurement of Copy Number Variation

As association studies depend upon accurate genotyping, the development of accurate CNV measurement techniques is absolutely vital (D. G. Clayton et al, 2005). However, the development of systems high on precision and accuracy has been problematic due to the complex structural nature of CNVs and the consequent genotyping challenge involved (McCarroll et al, 2008; McCarroll et al, 2007). Nonetheless, various methods of CNV measurement are available at present, but further development remains equally important.

1.3.1: Chromosome Analysis

Chromosome analysis or karyotyping corresponds to the study of the number and structure of human chromosomes (Tjio et al, 1956; Ford et al, 1956) and is accomplished through the analysis of bands created by banding techniques (Vogel et al, 1997). In a routine chromosome analysis, metaphase slides are prepared from actively dividing cells which have been arrested in the M phase of mitosis by the addition of a mitotic inhibitor, colchicine. These cells are then subjected to a hypotonic environment. Due to osmosis, the cells swell and after the addition of a fixative, they are permanently set onto the slide (Hsu et al, 1979). After trypsinization and staining with a specific dye (Craig et al, 1993), a characteristic banding pattern is visualized via a microscope. This characteristic banding pattern is seen for every chromosome pair which aids in its identification and correlation with specific sizes (Caspersson et al, 1968).

Regions stained as dark G (Geimsa stained) bands undergo late replication, contain more condensed chromatin and are less active transcriptionally, while R bands (Light G bands) undergo early replication and have less condensed chromatin (Craig et al, 1993). Genes are mostly concentrated in R bands. Differences between G and R bands also arise due to the type of dispersed repeat elements present in them.

On the basis of the type of staining dye or technique used, characteristic banding patterns are observed which include:

G-Banding involves Giemsa staining. Dark bands which stain positively with Giemsa (DNA binding chemical dye) are G bands whereas the negatively stained bands appear pale.

Q-Banding involves dyes Quinacrine, DAPI (4',6-diamidino-2-phenylindole) which stain the AT-rich regions of DNA and highlight the same regions as G bands but are fluorescent dyes .

R-Banding constitute the reverse of G bands which includes either saline heat denaturation of AT rich DNA or highlighting GC rich DNA by using GC specific dyes like chromomycin, olivomycin or mithramycin and therefore the pattern produced is Q/G negative.

T-Banding identifies telomereric regions particularly smaller than the ones identified by R bands. T bands are visualized by Giemsa or fluorochromes subsequent to thermal denaturation.

C-Banding is indicative of constitutive heterochromatin present at the centromeres and the procedure includes denaturation through barium hydroxide saturated solution before staining with Giemsa.

Earlier banding patterns provided a crude estimate of structural variations. 1-10 Mb regions could be inferred by this method. However, with the advent of higher resolution banding 400, 550 and 850 bands can now be analysed. These are mostly associated with the usage of prometaphase elongated chromosomes instead of metaphase chromosomes.

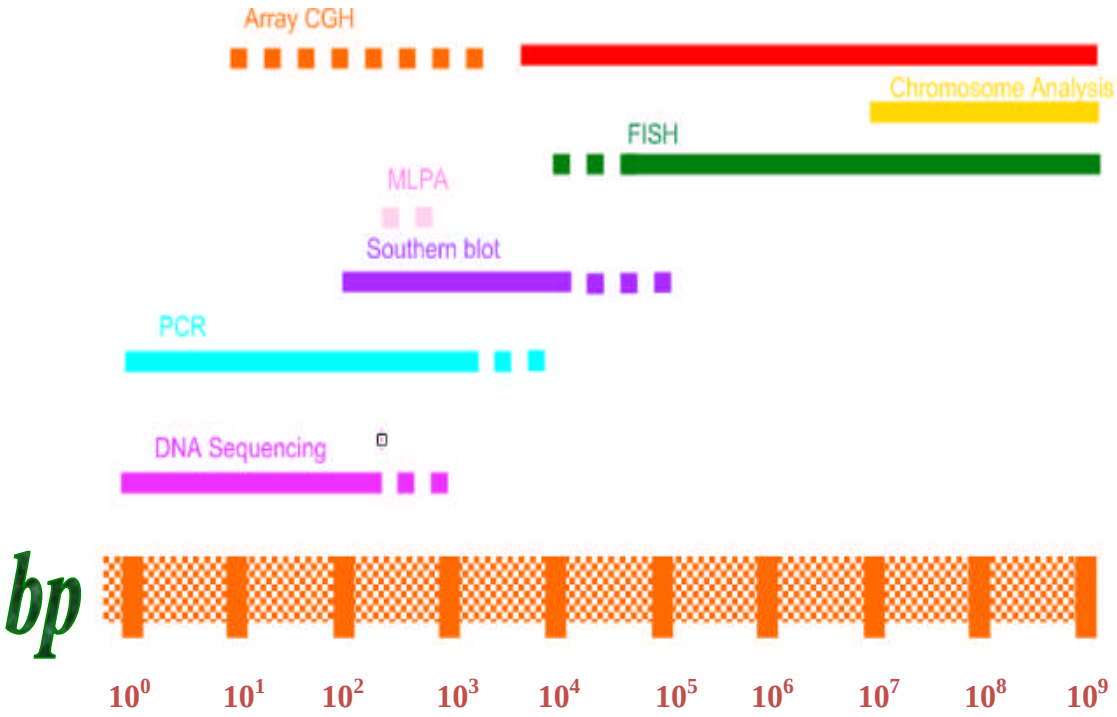


Figure 13. Resolution of different methods for measurement of structural variations.

1.3.2: Fluorescent in situ hybridization (FISH)

Fluorescent *in situ* hybridization is a molecular cytogenetic technique which allows detection and localization of a specific DNA sequence in interphase or metaphase chromosomes (Trask et al, 1991). The basic strategy behind FISH and chromosome analysis is the same except that instead of dyes, fluorescent probes are used which confers an added advantage of sequence specificity to the technique (Landegent et al, 1985*).*

Chromosomal preparations which contain cells arrested in metaphase or interphase for fibre FISH) are initially denatured in order to convert the dsDNA into ssDNA so that probes can bind efficiently to it. Probes containing the gene of interest or a part of the gene of interest and are cloned and hybridised to the denatured metaphase preparation (interphase in case of FIBRE FISH, van den Engh et al, 1992) so that they bind only with the specific chromosomal area which is complementary to it. Probes are labelled either with radioactive isotopes, immunogenic hapten conjugated nucleotide analogue or fluorescent dyes. Earlier, autoradiography was used to hybridize DNA with radioactive isotopes (Pardue & Gall et al 1969) but recently hapten labelled probes are used to detect the target DNA sites with the help of fluorescently conjugated antibodies.

Generally, the probe which is composed of DNA complementary to the target region of interest, is biotinylated and then hybridized to the denatured preparation and to detect the location of the hybridized probe, avidin or steptavidin conjugated with a fluorescent dye is added to the slide. This complex binds to the biotin of the hybridized probe. After the unbound avidin/streptavidin dye complexes are washed away, the slides are viewed with epiflourescence microscope armed with digital imaging. When excited by the light of appropriate wavelength, the bound dye fluoresces and the signal is captured electronically. Counterstaining with a different specific fluorochrome emits a different colour and can be used to show the chromosomal background to the signal from the probe. The chromosomal location of the probe is determined by identifying the complementary probe regions on the homologous chromosomes through matching the sites of fluorescent spots with chromosome banding pattern of the photographed metaphase spread. Resolution achieved by a typical FISH procedure is generally more than 1 Mb, and when prometaphase chromosomes are considered then problems relating to chromatin folding often confuse the analysis process. In

case of Fibre FISH, interphase chromosomes are chosen as they are less condensed and can be stretched along the glass slide during chromosomal preparation (Wiegant et al, 1992).

1.3.3: Comparative Genomic Hybridization (CGH)

CGH is also termed the 2 colour FISH Procedure. Principally, CGH is based on the co-hybridization of two differentially labelled DNAs (for example, tumour and normal) to human metaphase spreads and allows detection of DNA sequence copy number changes throughout the genome in a single hybridization reaction (Kallioniemi et al, 1992).

Briefly, in case of CGH, the total genomic DNA content is isolated from a test source and a reference source, one is labelled with green fluorescent dye (biotin dUTP) and the other with red fluorescent dye (dioxigenin dUTP). These two DNA samples are combined along with an unlabelled sample of highly repetitive DNA to normal metaphase preparations (Kallioniemi et al,1993). The unlabelled repetitive DNA competes with the labelled repetitive DNA from binding to the chromosome DNA and as a result only labelled single copy DNA sequences hybridize to the chromosome DNA. The relative intensities of fluorescence between test (Green) and reference cells (red) indicate the loss or gain of DNA sequences. If DNA is lost from the chromosomal sub region in test cells, then the comparable regions of the hybridized chromosome will fluoresce with more red than green.

Computer-assisted image analysis distinguishes between intensities of different fluorescence and analyses the relative amount of each fluorescent dye in defined segments of chromosomes and depending upon whether the green/red ratio is greater or less than 1, analysis of gain or loss is identified (Kallioniemi et al,1993). Chromosomal CGH has a resolution of 10-20 Mb and hence smaller variations are not detected. For example, when pancreatic carcinomas were screened by CGH, gain of sequences were consistently found over the chromosomal regions of 3q,5q,7q,8q,12p and 20q and losses in 8p,9p,17p,18q,19p and 21(Ghadimi et al, 1999).

1.3.4: Array CGH

Array CGH is an improvement of conventional CGH where instead of the metaphase spread which was hybridized to the test and reference samples, array CGH has arrays where target

DNA sequences (probes) are spotted or directly synthesised onto the slide (Pinkel et al., 1998).The principle behind both is the same and relies on the hybridization between differentially labelled DNA samples. In CGH, the exact sequence and position of every probe on the chip is known before hand and therefore any nucleotide fragment that hybridizes to a probe on the array can be identified by means of its genomic coordinates. Probes could represent a specific region of interest or entire genome. The number of probes varies between different commercially available platforms and hence the resolution depends upon the density of platforms. However, the number and size of the probes are not the only parameters that influence the resolution; genomic spacing and hybridization sensitivity also influence resolution.

There are three basic types of array platforms available for copy number analysis

1.3.4.1: BAC Arrays

Genome-wide array is generated by spotting locus specific BACs onto the array. Resolution is dependent upon the size of the BAC clone and also the spacing between them. If yields are low then DOP-PCR is used to increase its yield. DOP PCR (degenerate oligonucleotide primed PCR) utilizes partly degenerate oligonucleotide primers to amplify related segments of target DNA so that closely related segments are amplified simultaneously.

1.3.4.2: Oligonucleotide Array (Affymetrix Chips)

Locus specific oligonucleotide probes especially customized to match the genomic region of interest and often restricted by the genome architecture, particularly the repeat elements. They are synthesised directly onto the array slide (Brennan et al., 2004). In principle, test and reference DNAs are labelled with different flourophores (green and red), mixed and hybridized to the array. Since the two differently labelled DNA samples will compete in order to hybridize with the array, this competition will serve to provide differentially coloured spots and ascertain gain or loss of those sequences (Carvalho et al, 2004). The presence of a deletion is indicated by the representative spots as more red and if insertion occurs then more green is seen. The array is scanned and the red/green ratio is calculated by the computer software. Ratio values for the genes of interest are plotted onto chromosome

ideograms based on their mapping position which results in high resolution mapping of specific genomic imbalances. Resolution of oligonucleotide array varies between 30/50 Kb.

1.3.4.3: SNP Array

SNP arrays utilize oligonucleotide probes identifying allelic variants of specific SNPs (Zhao et al, 2004). Hybridization of genomic DNA to both probe variants is reflective of heterozygosity and deletion of single allele is suggestive of homozygosity (Zhou et al, 2004). Copy number is eventually deduced from the strength of the fluorescent signal emitted from an individual probe. However, since it relies on the allelic variants of SNPs this technique is not considered accurate because of the fact that SNPs are not evenly distributed over the entire genome and moreover some regions of the genome are not represented at all (Locke et al, 2006)

1.3.5: PCR Based Methods

Due to differential amplification dynamics, PCR techniques have not been considered quantitative but various alterations to the traditional PCR methodology has made it beneficial especially for eliciting quantitative information.

1.3.5.1: Real-Time PCR

As opposed to normal PCR where detection occurs at the end, real-time PCR relies on simultaneous detection and quantitation (Higuchi et al, 1992;1993). Real-time PCR reactions are defined by the period of time (cycle number) in the amplification process when the amplified target is first detected (exponential phase) rather than by the amount of product accumulated. Samples exhibiting a higher copy number of target genes are detected and amplified earlier as compared to the samples with lower amount of target genes which would eventually take more cycles to be detected. This occurs because the product detection only occurs above a threshold amount which is attained first by the samples with higher number of target genes.

Basically, there are different types of assays which can be used for quantitative PCR as discussed below.

TaqMan probes consist of oligonucleotides having a 5’ fluorescent dye and a 3’ quencher moiety. Initially the dye and the quencher molecules are in close proximity; therefore the fluorescence signal is not released from the probe. Gradually during the course of reaction, polymerase cleaves the probe due to its 5’ exonuclease activity and releases the fluorescent dye molecules. With each cycle, the amount of fluorescence increases which is proportional to the amount of probe cleavage and hence to the amount of product made (Heid et al, 1996;Jung et al, 2000).

Molecular Beacons exhibit a stem loop structure in the solution, with a flourophore attached to its 5’ end and a quencher at its 3’ end. In contrast to the taqman probes, molecular beacons remain intact during amplification and it is only due to the target hybridization occurring at each cycle that both flourophore and quencher get split and detection occurs.

The basic structure for scorpion probes is nearly the same as a molecular beacon, both contain a 5’ flourophore and a 3’ quencher, but scorpion probes also contain a sequence at the 3’ end which is complementary to the sequences present in the extension product of the primer. This special sequence is connected to the 5’ primers through non-amplifiable monomers (blocking). Thus after extension of a target sequence, probe hybridization occurs by means of a 3’ complementary sequence which leads to the release of fluorophore from the quencher molecule.

SYBR green includes a non-specific method of detection which binds to double stranded DNA and leads to a release of fluorescence when excited with light. However, problems due to overestimation of targets is often encountered with this method as SYBR green binds to any double stranded DNA including primer dimers.

As the amount of DNA product theoretically doubles with each PCR cycle, 2^n represents the amount of DNA accumulated after n^{th} cycles. Since PCR product increase within the initial few cycles would not be seen clearly on a linear scale, a logarithmic scale is used as the changes are much more pronounced.

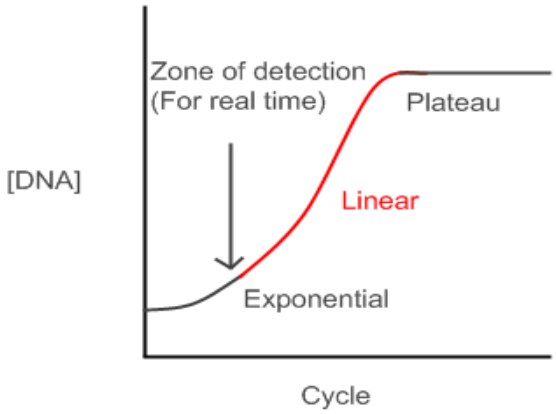


Figure 14. Relevant quantitative phases in PCR.
Optimal point for quantitative data collection is the late exponential phase. End point detection necessitates data collection at a fixed cycle number (variable signal)

Typically, quantification of real-time PCR is done through the comparative C_T (threshold cycle) method which can be understood through the PCR phases. C_T values relate to the accumulation of product and are thus related quantitatively to the initial input DNA. For example lower C_T values correspond to a greater amount of starting template, since fewer cycles are needed in order to detect the product. During the exponential phase where components are not rate limiting C_T values are most precise, which is the basis for real time quantitation by the use of standard curve analysis (Higuchi et al, 1992). As the name suggests comparative C_T method involves comparison of C_T values between samples of interest and the control samples (calibrator). Theoretically, it is assumed that the amplification efficiencies of both are the same and even upon dilution, difference between target and reference will always be the same, in other words, delta C_T will always remain constant. However, empirically it changes with dilution, in other words, delta C_T changes even with a 10% difference in PCR efficiency.

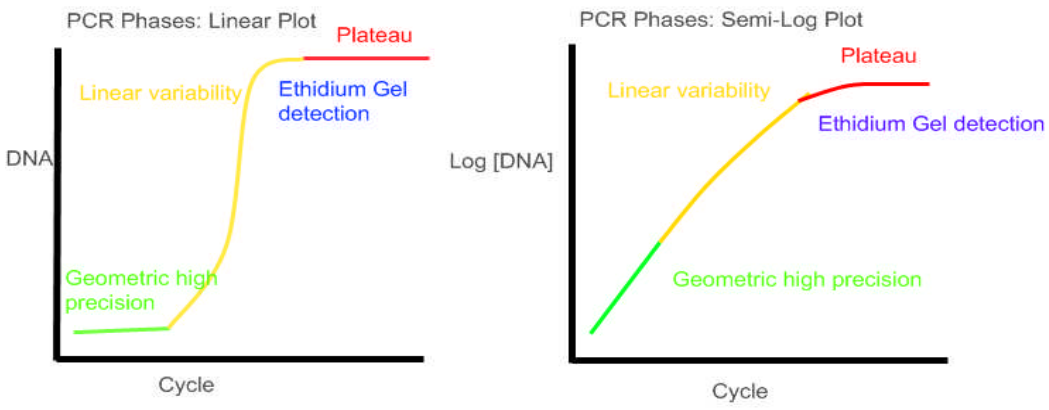


Figure 15. The linear and semi-log plots of a PCR cycle.
In the normal linear plot, minute changes occurring in the first few cycles cannot be seen. The C_T value is the point at which the fluorescence crosses the threshold.

At 100% efficiency, doubling of DNA occurs at each cycle. For 90% efficiency, the increase in DNA is 1.9 times; similarly for 80% efficiency, the DNA increase is 1.8 fold.

Efficiency	DNA increase per cycle	5 Cycles	10 cycles	20 cycles
100%	2	32	1024	1048576
90%	1.9	24.76099	613.1066	375899.7
80%	1.8	18.89568	357.0467	127482.4
70%	1.7	14.19857	201.5994	40642.31
60%	1.6	10.48576	109.9512	12089.26

Table 2. DNA molecules produced per cycle influence the efficiency of the PCR.
With decreasing efficiency, less amount of DNA is produced in each cycle.

1.3.5.2: Paralogue Ratio Test (PRT)

Developed by Armour *et al* (2007), the Paralogue Ratio Test utilises a single pair of primers, one of which is labelled, to amplify copies of both a copy number variable gene denoted "test" and a reference locus which does not vary in copy number. These products which differ in size are then separated and quantified by capillary electrophoresis, yielding a test to reference ratio. Since this method involves measuring a copy number variable gene with respect to its copy constant (2 copy) paralogue, it is known as the Paralogue Ratio Test (PRT). Utilizing a single pair of primers to amplify both a variable repeat unit and an unlinked locus prevents the inaccuracies caused in case of using different primers to amplify test and reference due to the relative amplification efficiencies of test and reference.

Apart from precision, particularly while interpreting the copy number of high copy number samples, PRT is both robust and inexpensive, (requiring only 10ng of genomic DNA), high throughput and ensures rapid copy number typing of large cohorts of samples in association studies.

Since PRT systems measure copy number as a diploid state, it is unable to ascertain the haplotype combinations of the copy numbers determined when used alone. Moreover, PRT systems would not be suitable for measuring balanced variants like translocations and requires other assays like indel measurement assays and microsatellite analyses to verify and validate the measured copy number. Many PRT systems have been designed and published, including those discussed below.

1.3.5.2.1: PRT systems for CCL3L1

CCL3L1 is located on chromosome 17q12, and is a multiallelic copy variable region, with individuals generally possessing 0-4 copies in Europe and up to 14 copies in Africa (Gonzalez *et al*., 2005). Therefore, effectively distinguishing copy number states within this higher copy number range is often challenging particularly when studies have shown that higher copy number of *CCL3L1* is correlated with the risk of Rheumatoid Arthritis (McKinney et al., 2008). Lower copy number of *CCL3L1* has been shown to be associated with enhanced HIV susceptibility and deviation from the average copy number has been said to be associated with systemic Lupus Erythematosus risk (Mamtani et al., 2008).

The CCL3L1 locus is 90 kb in length and contains the genes CCL3L1 and CCL4L1 (Townson, Barcellos, & Nibbs, 2002) with each repeat unit flanked by copies of the gene TBC1D3. The region contains the paralogues CCL3 and CCL4 that are 2 copy-constant (1 copy per chromosome) and are very similar to CCL3L1 and CCL4L1 in terms of protein and nucleotide sequence.

3 PRT systems have been developed for this locus:
CCL3A which compares CCL3L1 with CCL3.
CCL4A which compares CCL4L1 with CCL4.
LTR61 which measures the long terminal repeat with unlinked reference to chromosome 10.

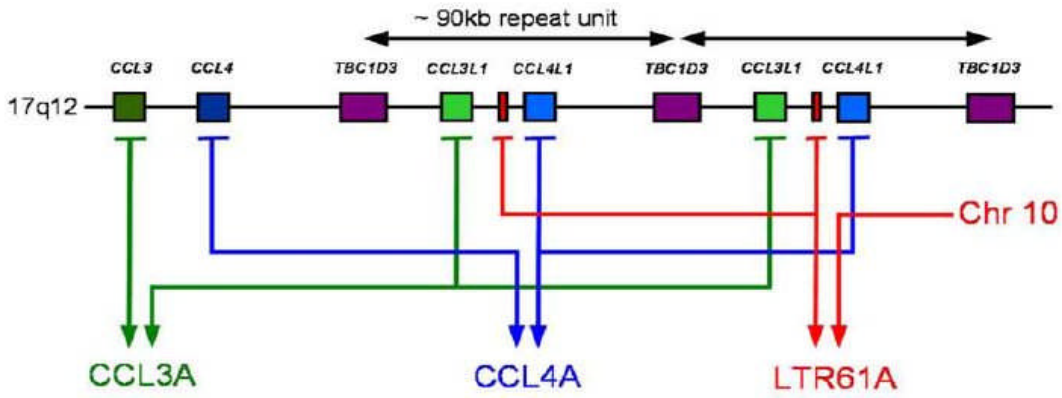


Figure 16. Genomic region representing the different PRT systems at CCL3L1 region. Source: Walker et al, 2009

As compared to real-time PCR based calculations of CCL3L1 copy number, PRT based systems are more accurate in large scale studies as calculations from real-time PCR results are affected by interference from 5' truncated CCL3L1 pseudo gene (Walker et al, 2009). Since this pseudogene lacks exon1 and the majority of intron 1, a PRT system was designed within intron 1 so as to avoid amplification of the pseudogene and subsequent inflation and overestimation of copy number (Walker et al, 2009).

1.3.5.2.2: PRT system for β-Defensins

The β defensin genes are located in three main gene clusters in the human genome, two on chromosome 20 and one on chromosome 8 (8p23.1). At the chromosomal locus 8p23.1, there is a variable repeat unit of at least 240kb, flanked by olfactory repeat regions. The genomic region comprises of seven defensin genes, including DEFB4, DEF103 and DEFB104 etc, which are polymorphic in copy number. Since the copy number of a copy-variable region varies as an entire unit, measuring each element within the copy number variable unit would essentially give the copy number for an entire unit.

In the case of β defensins, copy number usually varies between 2 to 7 copies per diploid genome (Armour et al, 2007), but may include as many as 12 copies and thus enhancing the possibility to identify euchromatic variants. These genes encode for antimicrobial peptides which are released by the immune system so as to fight against foreign pathogenic infections. A high copy number of this gene is associated with risk of Psoriasis (De cid et al, 2007).

As part of the development of PRT based systems, primers were designed to amplify HSPDP3, a pseudogene found 2kb upstream of DEFB4 within the copy number variable region at chromosome 8p23.1 and an unlinked reference sequence at chromosome 5, bearing similarity with the sequence at chromosome 8. Even though HSPDP3 is present in other parts of the genome, the primers made had mismatches against the other loci but a perfect match for the ones present at chromosome 8 and 5. This identity of both the sequences with the primers is very essential because substitutional mismatches between reference primer and reference sequence would cause drop out of single copy from the reference. This would lead to inaccurate doubling of the copy number of the test locus and hamper accurate PRT measurements. Inaccurate PRT measurements could also be due to rare variants, because of gene conversions among diverged gene families. Therefore it is essential to confirm the results by using a second PRT system with a different reference locus using a different part of the repeat element.

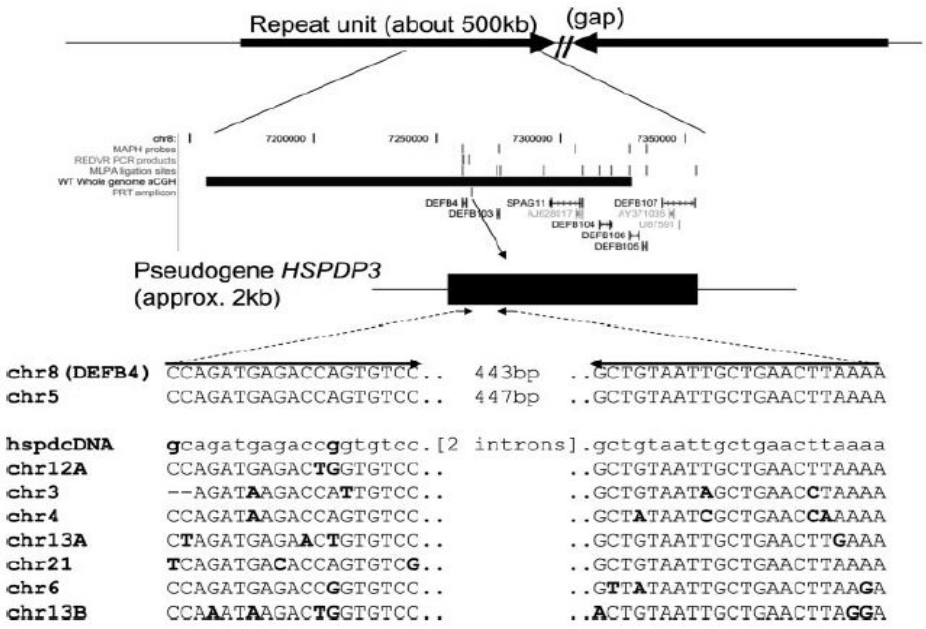


Figure 17. The Genomic region encompassing the β defensin genes at chromosome 8 with pseudogene and its perfectly matched primers at chromosome 8 and 5.
The primers designed also have mismatches to other sequences which could get amplified should the mismatches not been present.
Source Armour et al, 2007

Since the test (chromosome 8) and reference (chromosome 5) gave PCR products which were too close in size range (443 and 447 bp respectively), the products could not be resolved by capillary electrophoresis. Only after digestion with *Hae* III to give products of 302 and 315 bp could the products be distinguished. After *Hae* III digestion, products of HSPDP3 were known to give alternative fragment length and the absence of those fragments after digestion reinstated the fact that the products seen were exclusively from chromosome 5 and chromosome 8 and not from other loci.

1.3.5.2.3: PRT system for FCGR3

FCGR3 represents a multiallelic copy number variable region present on chromosome 1. FCGR3A and FCGR3B encode for two different isoforms of FCγRIII. FCγRIII is a cellular receptor of IgG and IgE. The isoform receptor encoded by FCGR3B is FCγRIIIb, which is a variant form of FCγRIIIa produced by mutation from arginine to stop codon leading to a truncated product. These two isoforms also differ in their mode of attachment and expression pattern . FCγRIIIa possesses a transmembrane region for attachment and is mostly associated with natural killer cells, whereas FCγRIIIb has a cell membrane glycophosphoinositol anchor and is mainly associated with neutrophils (Ravetch & Perussia, 1989). FCGR3B is correlated with FCγRIIIb protein levels and has been found to be associated with systemic autoimmune diseases like systemic lupus erythematosus (Aitman et al., 2006). Interestingly, FCGR3B duplication has been a characteristic feature of Asian population (Japanese and Chinese) ref. The most common copy number by PRT analysis is 4 (two copies of FCGR3A and 2 copies of FCGR3B), although the range of copy number found was 2-7 copies per diploid genome (Hollox et al, 2008).

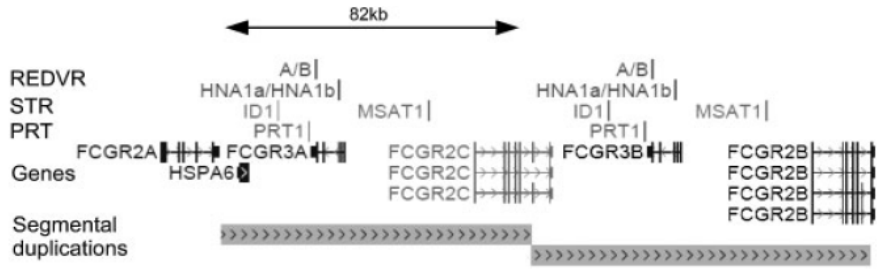


Figure 18. The genomic region occupied by FCGR3 on chromosome 1q23.3.

A single set of primers were used to amplify two paralogous repeat regions (82 kb) present on chromosome 1 along with one paralogous repeat region present on chromosome 18. Two amplifications (HEX and FAM) per sample were carried out. Later based on the size differences of test (67bp) and reference (72bp), they were separated by capillary electrophoresis (Hollox et al, 2009).

1.3.5.3: Multiplex Amplification Probe Hybridization (MAPH)

The basic strategy behind MAPH involves hybridization of probe DNA sequences with denatured input genomic DNA immobilized on a small nylon membrane (Armour et al, 2000). Excess probes are used to ensure complete annealing of homologous genomic sequences thus driving hybridization. Washing ensures that unbound probes are fully removed. Afterwards the bound probes are simultaneously released and quantitatively amplified with universal primers. The recovery of the bound probes is a critical step in accurate measurement of copy number through MAPH as the amplification of these bound probes are indicative of the copy number of that particular locus. Bound probes contain certain signature sequences which act as markers for quantification due to the fact that it exhibits complementary base pairing with the specific regions of interest therefore quantifying probes would imply indirect assessment of the particular genomic region in question. Amplification products are visualized on polyacrylamide gels or capillary electrophoresis and each band is quantified against other probes with known copy number. Probe selection is also a critical step in accurate measurement of copy number by MAPH. The sequences selected for designing probes are unique, non repetitive, non polymorphic and having similar GC content.

Two different methodologies are used for the development of probes:-

1. Identifying the specific DNA sequences in the region of interest bioinformatically and after its amplification with unique primers, sub cloning it (amplified product or probe) into a plasmid vector so that numerous probes can be generated.
2. Genomic clones can be made to undergo a blunt end digestion. After digestion, they are inserted into the *EcoRV* site of pZero 2.

All probes which are subcloned in the same cloning vector are flanked by same sequences (universal primer sequences) which can be amplified simultaneously in a single reaction. Its resolution varies from 100-300 bps.

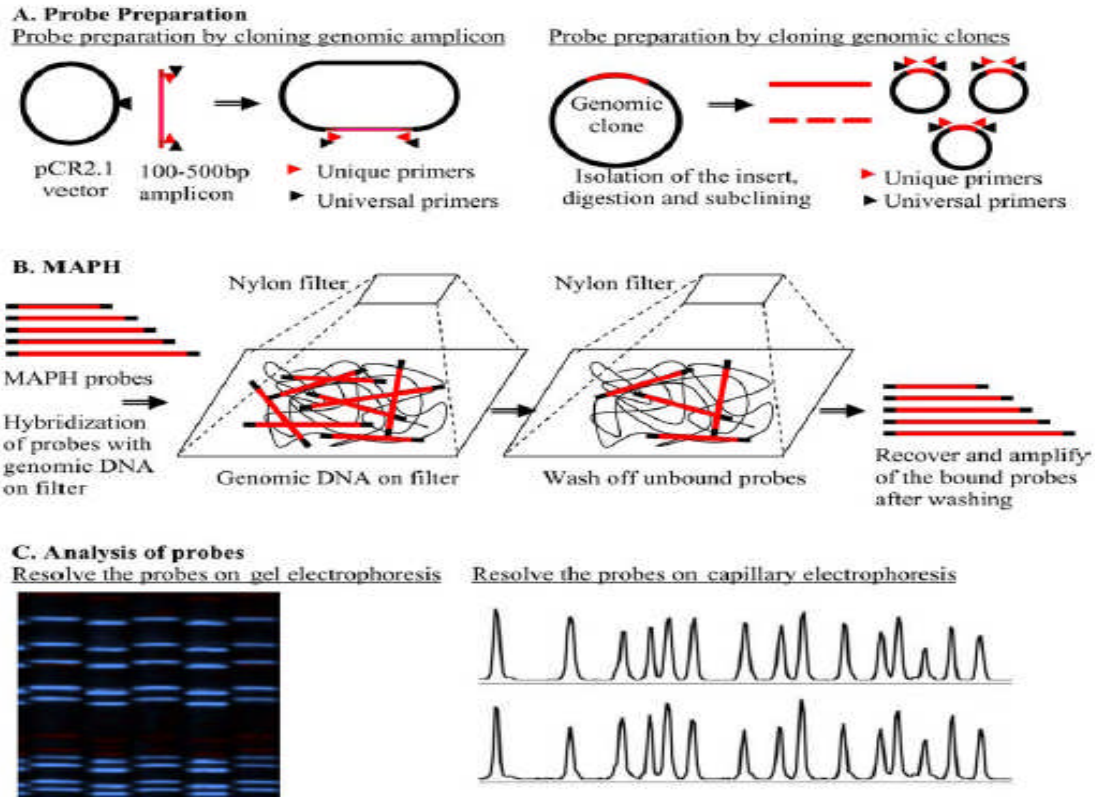


Figure 19. Multiplex amplification probe hybridization procedure whereby quantification of the recovered hybridized probes is used to ascertain variation in copy number.
Source, Patsalis et al, 2005.

1.3.5.4 : Multiplex ligation-dependant probe amplification (MLPA)

Like MAPH, MLPA also relies on the enumeration of copy number through quantification of amplified probes that indirectly assesses the complementarily bound sequences as well (Schouten et al, 2002). Unlike, MAPH, MLPA uses half probes and requires only 100-200 ng of input DNA. Each probe has two halves- 5' probe and 3' probe. These probes contain the sequences that are complementary to the genomic region of interest. Each half probe (apart from the target sequence) also has a universal primer sequence along with a stuffer sequence which may be present in one or both half probes. This stuffer sequence is usually of variable length which aids in distinct separation of different targets. In addition the 3' half probe is 5'

phosphorylated. This is done to ensure that ligation between the 2 half probes and subsequent amplification only occurs if and when the entire probe (with both halves) is hybridized to the target locus. Subsequent PCR generates loci specific amplicons which differ in size (due to stuffer sequences) and can be resolved by capillary electrophoresis. MLPA has a resolution of ~60 nucleotides and hence is able to detect variations within a single exon.

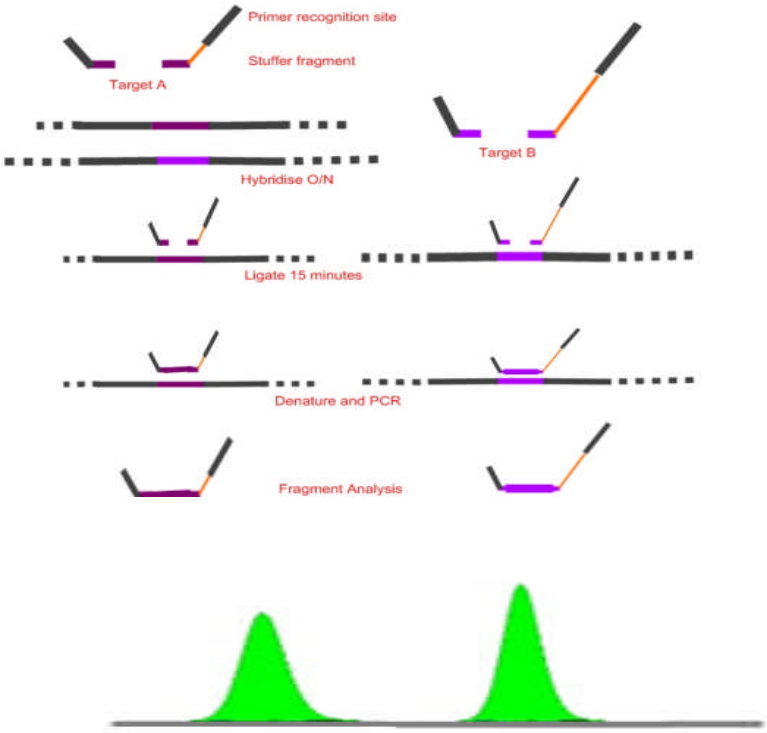


Figure 20. The multiplex ligation dependant probe ligation procedure. Detection of copy numbers variation is made easier by the incorporation of stutter sequence which increases the size difference between similar sized amplicons.

1.4: Salivary Amylase

α Amylase (α-1, 4- glucanohydrolase) is an enzyme which initiates catabolism of α-1, 4-glycosidic bonds present in starch and glycogen to maltose. In humans, amylase (ptyalin) is secreted by exocrine glands, primarily the salivary glands and pancreas (Merritt et al, 1973) and aid in the digestion of starch which is composed of amylose (linear polymer of D Glucose units insoluble in water) and amylopectin (branched glucose polymer soluble in water). Digestion of starch initially begins in saliva, where insoluble molecules of starch (amylase) are converted to soluble starches (amylodextrin, erythrodextrin and achrodextrin) and is continued by pancreatic amylase where the digestion of soluble starches (amylopectin)

occurs to yield dextrin, maltose and maltotriose). Later thse soluble starches are broken down by enzyme maltase or α dextrinase into glucose inside enterocytes residing in intestinal microvilli.

Though the human amylases were initially thought to be coded by two different genes (Merritt and Karn, 1977) because of differential mobility and antigenicity, it is now known that the human amylases are encoded by two paralogous genes AMY1 (Samuelson et al, 1988; Perry et al, 2007), responsible for secretion of salivary amylases and AMY2 responsible for secretion of pancreatic amylases. The tissue-specific expression of these two genes represents a very unusual feature of this locus in spite of sharing sequence similarity with each other.

Studies involving salivary and pancreatic cDNA sequence analysis revealed the extent of similarity between the two. Coding regions of AMY1 and AMY2 corresponded to 1768 and 1566 nucleotides respectively. Sequence identity exists between the non-coding regions of AMY1 which includes 200bp in 5' UTR and 32 bp in 3'UTR and the non coding regions of AMY2 which includes 3bp in 5'UTR and 27bp 3' UTR, which accounts for 96% nucleotide sequence identity between AMY1 and AMY2 in the coding region and predicted amino acid sequence identity of 94% (Nishide et al, 1986).

Studies by three independent groups provided insights into the complicated nature of this particular gene cluster. In one particular study, human genomic DNA cosmid library was screened by mouse amylase cDNA and the resultant 230kb fragment was found to contain five complete amylase genes and one truncated pseudogene; ribonuclease protection assay validated strict tissue specificity (Samuelson et al, 1988). Another group generated overlapping cosmid clones from a human genomic library screened by a human pancreatic amylase cDNA (Groot et al, 1989, 1991). Although there was some disagreement between these studies, they agreed on the basic structure of the locus represented by two pancreatic (AMY2A and AMY2B) and three salivary amylase genes (AMY1A, AMY1B, AMY1C). It was also reported that the three salivary amylase genes were a product of very recent duplication.

Another study focussed on amylase transcriptional analysis which revealed that the amylase gene is about 10kb in length, with 11 exons and 10 introns (Nishide et al, 1986). It was later

revealed that even though the structures of salivary and pancreatic amylases exhibited high sequence similarity to each other, the presence of an untranslated exon at the 5' end of salivary amylase sets them apart (Horii et al, 1987). CA microsatellite repeats were found to be located 1kb upstream of the amylase gene cluster (Gumucio et al, 1986) and these dinucleotide alleles were observed to vary in number between 16 to 21 (Dracopoli and Meisler et al, 1990). According to somatic cell studies (Munke et al, 1984, Tricoli and Shows et al, 1984) and *in situ* hybridization studies (Zabel et al, 1983), the amylase cluster was mapped to the chromosomal band 1p21.

Through Southern blot analysis of hybridization between amylase cDNA probes and human genomic DNA, quantitative variation in terms of copy number for amylase was detected (Groot et al, 1989,1991). Sequence comparison of the proximal promoter regions of the amylase genes have revealed the existence of two elements inserted into the promoter regions of amylase genes.

1. A 2kb processed pseudogene upstream of AMY2B (Samuelson et al, 1988) sharing 89% sequence identity with mRNA of human γ actin, was found to be in agreement with a divergence time of 40 million years (Samuelson et al, 1990). These signature sequences are present in all amylase genes indicating derivation from the same precursor ancestral gene in which other sequences got inserted (Emi et al, 1988; Samuelson et al, 1988, 1990).

2. The pseudogene is itself disrupted by the presence of retroviral elements upstream of the promoter regions of all Amylase genes excluding AMY2B (Emi et al, 1988, Samuelson et al, 1988).

Therefore all copies of salivary amylase genes are associated with an intact retroviral element and one of the pancreatic genes (AMY2A) is associated with sequences which are similar to some but not all of the sequences of the retroviral element which further suggests that maybe this particular pancreatic gene has been derived by the excision of the retroviral element gained by the salivary genes. Moreover, since transcription for AMY2 starts at exon a and the transcription of AMY1A starts from the nontranslated exon inherent to the pseudogene (Nishide et al 1986, Emi et al 1988, Samuelson et al 1988) therefore it is very likely that the

insertion of the retroviral element inside the pseudogene triggered the activation of a hidden promoter which assists in transcription of the salivary amylase locus.

For assessing the functional role of salivary amylase, AMY1C gene cosmids were used to make transgenic mice (Ting et al, 1992) and human amylase tissue specific was observed in parotid glands of transgenic mice. Further experiments involving deletion constructs and fusion genes ascertained the 1kb fragment embedded in the promoter region was responsible for conferring salivary specific determinant of tissue specificity. Apart from the parotid glands, salivary amylases in humans are also secreted by submandibular and sublingual glands (Whitten et al, 1988; Korsrud and Brandzaeg, 1982). Interestingly, some lung tumors and ovarian tumors (Zakowski et al, 1984; Hayashi et al, 1986, Tomita et al, 1988), thryroid and thyroid adenomas (Doi et al,1991) also produce amylase. Both parotid and submandibular gland produce amylase in rat (Shear et al, 1973; Bloom et al, 1975), but only parotid glands are associated with the production of amylase in mice (Hjorth et al, 1979).

Since transgenic mice exhibited high expression of salivary amylase in parotid gland, low expression in ovary and lung and no expression in submandibular gland, it is very likely that transgenic expression mimics amylase expression in humans with the exception of expression in submandibular gland, which can be explained on the basis of mouse submandibular gland losing its regulatory element after its divergence from a common ancestral gene or humans gaining this particular regulatory element after their divergence.

Further investigation revealed that the mouse genome also contains both salivary and pancreatic amylase genes (Schibler et al 1982; Wiebauer et al 1985). Insertion of a retroviral element into the proximal promoter regions of salivary amylase gene led to its divergence from the ancestral gene copy. Since this retroviral insertion occurred subsequent to its divergence, some regulatory features in the human and mouse genomes have evolved independently. This theory is supported by the conserved transcript structure of mouse and human genomes and absence of similar regulatory elements in mouse.

The convergent evolution in both mouse as well as human genome has been indicative of a strong positive selection for salivary amylase during mammalian evolution. It suggests that this particular locus is conferring some kind of selective advantage to the organisms possessing it. Expanding on this conjecture, further studies investigated the details of γ actin

and retroviral insertions particularly with regard to primate evolution by examining these sequences in extant primates.

Considering the fact that amylase gene copies of the old world monkeys were not found linked to a retroviral element, it was established that both these insertions occurred subsequent to the division of primates into new and old world monkeys. γ-actin pseudogene insertion occurred after divergence of new world monkeys and insertion of retrovirus occurred after separation of old world monkeys, that is somewhere after the orangutan separation from chimp-hominoid family (Goodman et al, 1986) as all primates analysed had pseudogene insertion but only the hominid lineage including humans had both γ actin and endogenous retroviral insertion (Samuelson et al, 1996).

Salivary amylase gene promoters of old world monkeys did not contain the retrovirus which suggests alternative mechanisms of regulation of salivary amylase must be existent within old world monkeys. New world monkeys did not contain the pseudogene, indicating the role of γ actin pseudogene in activation of salivary amylase from ancestral pancreatic amylase.

Research conducted by Perry et al (2007) focussed further on this theme and successfully correlated the effect of diet with salivary amylase gene expression and copy number in different populations; choosing populations not related to one another so as to eliminate the confounding genetic effects. The dietary intake of starch was a major determinant for selection of these populations especially for this particular study and upon comparison with chimpanzees, it was observed that while amylase diploid copy number varied up to 14 between different human populations, chimpanzees exhibited a copy number of only 2 which presumably reflects members of ancestral state of this locus.

Further comparisons suggested the possibility of the gain of copy number in human lineage according to dietary starch pressures and this hypothesis is consistent with the observation that new world monkeys are devoid of salivary amylase presumably because they consume little starch. The fact that in certain cercopithecines (old world monkeys) production of salivary amylase is higher than humans even though they are omnivorous further complicates the matter. As cercopithecines are characterised by cheek pouches, there are some theories which propose that higher amylase content in cercopithecines aids in digestion of foods temporarily kept in the cheek pouch by these organisms (Mau et al, 2010).

It is very likely that during the course of evolution, natural selection favoured the multiplication of the salivary amylase gene copies in humans particularly during the era involving a shift from hunter gathering to farming (consistent with chimpanzees having low but not complete absence of salivary amylase machinery which was enough to digest lesser amounts of starch). Gradually, variation with regard to amylase occurred within human population because of the differential intake of starch.

1.4.1: Aims

Development of high throughput and accurate typing systems for the determination of human amylase copy number variation by PRT ratios and microsatellite repeats.

1.4.2: Measuring Salivary Amylase Copy number

The human salivary amylase gene copies (AMY1) are mainly distinguished from the pancreatic amylase gene copies (AMY2) by the integration of a human endogenous retrovirus (ERV) into the upstream promoter regions of AMY1 gene copies. For the determination of copy number, the existence of these ERV insertions aids in the PRT construction which is designed by comparing the PCR yield from amylase associated ERVs with the ERV reference loci located elsewhere in the genome; but unlike most parts of the gene sequence, a clear distinction is made by which only salivary amylase copies (AMY1) are counted.

1.4.1.1: PRT 12A

PRT 12A is a copy number variable test associated with ERVs present on chromosome 1 and fixed two copy reference present on chromosome 12. After amplification, test and reference loci are resolved on the capillary based on their size differences (test 244 bp and reference 249 bp). Depending upon the test to reference ratio, copy numbers were calculated.

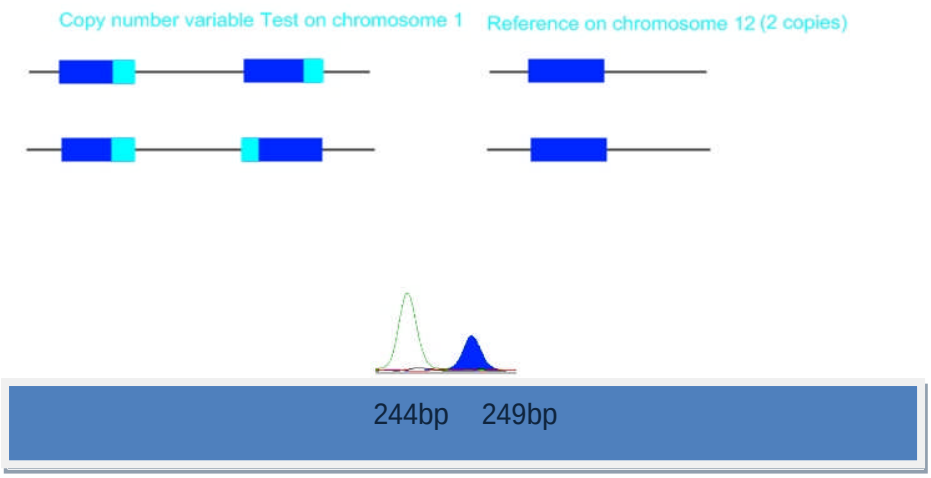


Figure 21. The strategy of 12A PRT design.
The dark blue boxes represent ERV paralogous sequences used to design this PRT system. As this ERV is associated with all AMY1 copies, quantification of ERV would indirectly measure copies of AMY1 (irrespective of its orientation) which would otherwise be difficult to measure considering the amount of similarity between AMY1 and AMY2 copies. One pair of primers are utilized to amplify the copy number variable region with respect to the copy number constant (2 copies) reference in order to avoid the difference between test and reference amplification efficiencies.

1.4.1.2: PRT 1H

PRT 1H is a copy number variable test present on the same chromosome (chromosome 1) as the reference. However, there is considerable distance between the test and reference sequences. After amplification they are quantified on the capillary based on their size differences (test 122 bp and reference 141 bp). Depending upon the test to reference ratio, copy numbers were calculated.

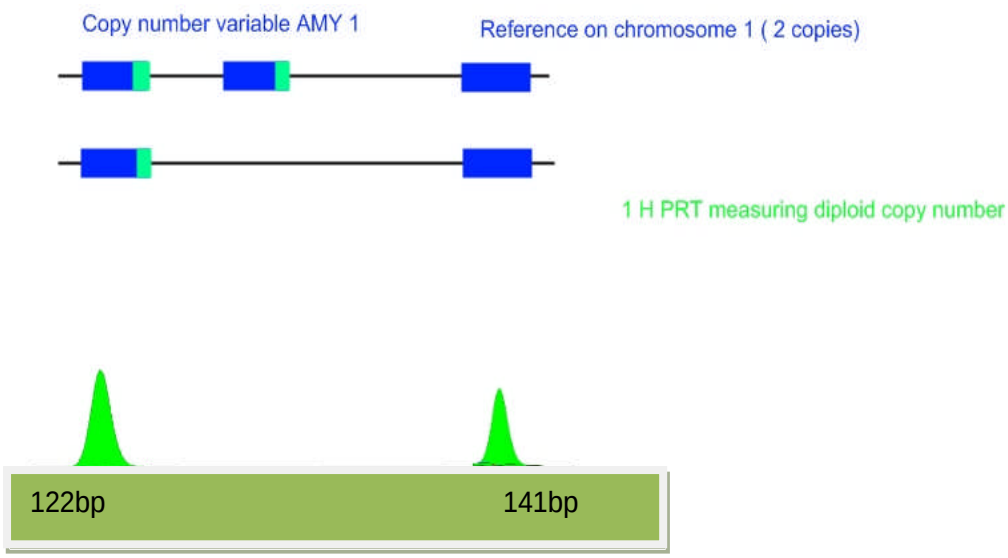


Figure 22. The design of 1H PRT where test and reference are both on the same chromosome.

CHAPTER 2: MATERIALS AND METHODS

2.1: Materials

2.1.1: Samples

Standard cell line and ECACC Human Random Control Panel 1 samples were chosen for most of the developmental part of the work. However, copy number calibration was carried out with HapMap Japanese samples (The International HapMap Consortium, 2003). These samples were chosen for referencing as their copy number was already known (Perry et al, 2007). Samples had DNA concentration of 10ng/µl.

2.1.2: Primer Designing

UCSC Genome Browser March 2006 Assembly was used to design primers by searching sequences in the target regions of interest. PRT primers were designed by means of a thorough investigation of the test target retroviral elements located upstream of AMY 1. The 8 kb retroviral element sequence was divided into 1kb fragments which was essential (limitation of PPPP, see below) BLAT was used to determine matching sequences located elsewhere in the genome. The results generated were then converted to a “bed” formatted file and minor reformatting was done to ensure the deletion of spaces and returns as the programme designed to find test (within the ~8 kb sequence) and reference (other regions) interprets spaces and returns as the end of sequences. Thereafter Bed formatting was used to recover sequences in bulk and these sequences were aligned by Clustal W. A C++ programme, PRT primer picking programme (PPPP written by John Armour) was utilized for identifying primers for potential test and reference targets.

Primers for amplifying microsatellite were designed through Primer3. When compared to AMY2 sequences, only one mismatch was present within the reverse primer amplifying the microsatellite but the presence of enough mismatches at the 3’ end of the forward primer ensured that only AMY1 was amplified. Trace archive was used to check the presence of SNPs within the primer binding sites of all the primers.

2.1.3: 10X LD Mix

Low dNTP PCR mix contained 50mM Tris-HCl (pH8.8), 12.5Mm ammonium sulphate, 1.4mM magnesium chloride, 125µg/ml BSA, 7.5Mm 2-mercaotoethanol and 200mM of each dNTP.

2.1.4: 10X PCR Mix

Generally the final volume used for PCR was 10µl which included 1µl of 10X LD Mix, 1µl of forward primer (1µM) , 1µl of reverse primer (1µM), 0.1 µl of *Taq* polymerase (0.5 units), 1µl of input DNA (10ng/µl) and water was added in order to make the final volume. Primers were ordered from Fisher, and redissolved to a concentration of 100µM.

2.2: Methods

2.2.1: PCR

Products analysable by agarose gel were amplified using 35-37 cycles of 95˚C for 30 seconds, 60˚C for 30 seconds and 70˚C for 1 minute. The number of cycles used was dependant on the type of separation technique involved, whereas the exact temperature was dependent on individual systems primers and the products expected.

2.2.2: DNA Electrophoresis

2.2.2.1: Agarose gel Electrophoresis

Amplified products representing individual samples were mixed with 10X loading dye buffer comprising of 0.025% bromophenol blue dye, 40% sucrose, 2.5x TBE buffer. The mixed solution was then run on 2-3.2% (w/v) gel comprised of agarose, 0.5X TBE and 0.5Mg/ml ethidium bromide. Apart from the amplified samples with loading buffer, DNA ladder (Ladder, loading buffer, water) was also loaded as a size standard and both were run at 110V for 2 hours. Bands were visualized under UV through a gel doc system.

2.2.2.2: Capillary Electrophoresis

An ABI Genetic Analyser 3100 (Applied Biosystems, UK) was used for this study. One of the primers was labelled fluorescently and PCR was carried out using the standard mentioned conditions. Amplified PCR products (1-2µl) were mixed with 10 µl of HiDi formamide and 2 µl of ROX 500 size standard (Applied Biosystems, UK). Thereafter denaturation at 96˚C for 3 minutes ensured that single stranded DNA molecules get separated on POP-4 polymer at an injection time of 10 seconds (Applied Biosystems, UK). Analysis of the generated peaks was carried out through gene mapper software which was also used to transfer the data onto a spreadsheet.

2.2.3: Restriction analysis for 12A PRT system

2.2.3.1: RFLP for 12A PRT system

For assessing the contribution of potential alternative loci, PCR product restriction digestion was carried out. Initially, PCR using the standard conditions was carried out using fam labelled forward primer. 2µl of PCR product was digested with 0.5µl of *Taq* I (5 units), 1µl Buffer, 0.1µl BSA and 5.5µl of water. It was followed by incubation at 65˚C for 6 hours to allow complete digestion. For the analysis of digested product, 2.5µl of the digested product was visualized over the capillary at an injection time of 30 seconds.

2.2.3.2: RFLP for 1H PRT system

Similarly, amplified products (2µl) from 1H PRT were digested with 2µl of buffer, 1µl of *Alu*I (0.5U) and 5.5µl water. Following incubation at 37˚C for 3 hours, 2.5 µl of digested product was visualized over the capillary at an injection time of 30 seconds.

2.2.4: Sequencing for 1 H PRT system

Initially, products were amplified using the general PCR conditions. Prior to sequencing, these products were cleaned using CleanSeq (Beckman Coulter) according to the manufacturer's instructions. BigDye terminator kit V3.1 was used for sequencing. Two sequencing reactions representing each of the primers per sample were set up. The total reaction volume set up for each sample was 10 µl containing template DNA which was purified, 1µM primer, 1X sequencing buffer (50mM Tris HCl pH 9.0 and 2mM Magnesium

chloride), Big Dye Terminator containing 4 deoxynucleotides and four dideoxynucleotides, each labelled with different fluorescent dye. The sequenced products were sent to DBS Genomics for electrophoresis.

2.2.5: Paralogue Ratio Test (PRT)

2.2.5.1: 12A PRT system

The human salivary amylase genes posses a characteristic feature of retroviral insertions located upstream of its promoter regions which distinguishes salivary amylase from other genes. PRT 12A utilizes a pair of primers to simultaneously amplify amylase associated ERVs located on chromosome 1 and ERVs located elsewhere in the genome (chromosome 12). Products were amplified using PCR as described in the section using FAM labelled forward primer. Test and reference products were resolved on the capillary with an injection time of 10 seconds.

2.2.5.2: 1H PRT system

This system utilizes a pair of primers (HEX labelled forward primer) for amplifying two paralogous sequences- test and reference located on the same chromosome. Products were amplified using PCR as described in sections with a predenaturation of 5 minutes followed by 23 cycles of 95˚C for 30 seconds, 51˚C for 30 seconds and 70˚C for 1 minute and a chase reaction of 72 ˚C for 20 minutes. Because of considerable size difference between the test and reference amplicons, PCR products are directly resolved on the capillary at an injection time of 10 seconds.

Assay	Primer Sequence	Annealing temperature	Test size	Reference size
PRT 12A	F*CTTGTTTTATATTGTTTTCTATT R-AGAAAATAAGTACATCTGTCAAG	62.5	244	249
PRT 1H	F*CTTCTAAATCATAAGTTGATTT R-CTTGTTTTATATTGTTTTCTATT	51	122	141
TG MS	F*ATCTTTCCTGAAGTTTTTCAATG R-AGAGTGTCCACATAAAAGCTAACA	59	98	100

Table 3. The assays developed along with their primer sequence, annealing temperature, test and reference size resolved on the capillary.

2.2.5.3: Data analysis

Since the project involved development of PRT systems, both test and reference height and area were recorded to evaluate concordance and also to examine the best measure out of the two parameters. Height and area ratios were calculated and calibrated with known copy number samples. The relationship (least square linear regression) between ratios and copy number representing individual samples was used to evaluate the copy numbers of other samples.

2.2.6: TG Microsatellite Assay

Products were amplified using PCR conditions involving 25 cycles of 95˚C for 30 seconds, 59˚C for 30 seconds followed by 72˚C for 1 minute with an additional step of 72˚C for 20 minutes which ensured complete 3' dA addition. It was further resolved on capillary at an injection time of 15 seconds.

2.2.6.1: Data analysis

Since microsatellites suffer from non specific amplifications, incomplete dA nucleotide addition and slippage effects, the microsatellites profile was analysed assuming that slippage contribution towards individual alleles was the same.

CHAPTER 3: RESULTS

3.1: Primer Design

Sequences upstream of the salivary amylase gene includes the endogenous retroviral region were downloaded from UCSC genome browser March 2006 assembly and BLAT was utilized to identify all other sequences in the genome having maximum nucleotide base pair similarity with the endogenous retroviral sequences associated with all salivary amylase copies. BED formatting ensured that the sequences were recovered as an entire set. Further reformatting was done to confirm the omission of spaces as the C++ compiler used to run the programme (PRT primer picking programme) for identification of potential primers for amplifying test and reference sequences exclusively interprets spaces as the end of sequences, hence spaces and returns had to be omitted for the correct alignment of each sequence. After aligning the sequences using ClustalW, PPPP was made to scan through the aligned sequences in order to identify the best matched primer pair having maximum similarity with the retroviral sequences and sequences present elsewhere in the genome so that only one pair of primers amplify both salivary amylase copies and only one such sequences present elsewhere in the genome and no other locus is amplified. Apart from the primer sequence and its location on the chromosome the output file generated by PPPP also contained other parameters such as mismatch scores as specified by the input file. The mismatch score represents the extent of mismatch between the primers amplifying the two targets (test and reference) and the alternative amplification targets and was an important parameter for selection of individual PRT systems.

Thus, for a particular primer pair set the higher a mismatch score the lower chances of it amplifying alternative targets other than test and reference. Therefore, for the development of individual PRT systems, preference was given to systems with higher mismatch scores.

With 1 reference:

mismatch score			PCR length		primer 1			primer 2		
ref	mean	min	product	test	ref	position	sequence	position	sequence	
A.>chr12A	882.822	6	5296.93	244	249	1188-1209	ATCTAGTCCTTTTCTATCAATG	1440-1463	TCTTGACAGATGTACTTATTTTCT	
B.>chr5A	347.908	4	1391.63	293	291	328-348	CTTATGGCTATGCTTCATTTT	633-667	TACCATGTAACTGTTTTTGT	
C.>chr4A	541.46	4	2165.84	284	288	101-121	TCTTTCTTCTGTTCTAGCTAT	402-420	AGTCCTCTGTGTGGGAAGA	
D.>chr12B	499.153	5	2495.77	147	145	239-260	TCACTTCCTGGATTACATACTT	373-403	GATGTATGCACTGAAGGCAG	
E.>chr4A	415.204	5	2076.02	245	245	609-628	TTCAGTGTAAGATGGAGTG	914-932	GGCTGTGGGCTCAGGAG	
F.>chr1H	277.391	5	1386.95	123	142	818-840	GCTTCTAAATCATAAGTTGATTT	982-1005	AATAGAAAACAATATAAAACAAG	

With 2 refrences:

mismatch score			PCR length		primer 1			primer 2		
ref	mean	min	product	test	ref	position	sequence	position	sequence	
G.>chr19F	242.596	4	970.382	268	249	483-500	TTTCCTCATGCTTTGGCC	745-763	TCCAGTCCTTGATCCACAC	
H.>chr4B	614.843	5	3074.22	74	71	914-932	GGCTGTGGGCTCAGGAG	1081-1099	GTTGGTGCCATCTCCTGCC	
I.>chr2A	306.173	4	1224.69	388	380	205-226	GGACTCCAAGTGCCAGTTCCTT	595-617	CCAAGTGCACAATTTCTGTCCCT	
J.>chr5A	258.679	8	2069.43	359	351	238-259	GCCACTGTGTTAATCCTCCGCG	597-621	AAGTGCACAATTTCTGTCCCTTTTA	
K.>chr10K	258.679	8	2069.43	359	351	238-259	GCCACTGTGTTAATCCTCCGCG	597-621	AAGTGCACAATTTCTGTCCCTTTTA	
L.>chr1E	306.173	4	1224.69	388	380	205-226	GGACTCCAAGTGCCAGTTCCTT	595-617	CCAAGTGCACAATTTCTGTCCCT	
M.>chr5A	257.199	7	1800.39	357	349	240-261	CACTGTGTTAATCCTCCGCGGG	597-621	AAGTGCACAATTTCTGTCCCTTTTA	
N.>chr10K	258.189	7	1807.32	358	350	239-260	CCACTGTGTTAATCCTCCGCGG	597-621	AAGTGCACAATTTCTGTCCCTTTTA	

With 3 refrences:

mismatch score			PCR length		primer 1			primer 2		
ref	mean	min	product	test	ref	position	sequence	position	sequence	
P.>chr2A	277.718	4	1110.87	365	357	232-255	TGTTCAGCCACTGTGTTAATCCTC	597-621	AAGTGCACAATTTCTGTCCCTTTTA	
Q.>chr4C	128.569	4	514.277	363	355	237-258	AGCCACTGTGTTAATCCTCCGC	600-624	TGCACAATTTCTGTCCCTTTTAAGG	

Figure 23 The potential PRT system primers generated through PRT primer picking programme (PPPP).
This programme scans through the query sequences and identifies primers for coamplification of paralogous sequences (test and reference) with their position and mismatch scores. Mismatch scores evaluate the mismatches between the primers and alternative amplification targets (apart from test and reference sequences) so a higher mismatch score would indicate primer specificity for exclusively amplifying test and reference.

After identifying the potential primer results generated by PPPP, the NCBI human trace archive was used to check for polymorphisms within primer binding regions by pasting the individual BED derived sequences into trace archives. PCR based copy number measurement systems introduce error which is caused due to amplification failure of individual gene copies because of the presence of sequence variants within them.

Initially development of PRT systems for amylase genes with two or three reference loci was given preference because of their added precision for measurement, but due to the problems associated with their construction as explained below, they were discarded.

3.1.1: Identification of two reference systems for salivary amylase PRT construction.

Two reference systems include a primer pair that accurately amplifies the test (copy number variable) and two other such loci as a reference (fixed copies) anywhere else in the genome. All these systems were named according to the BLAT results specifying a chromosome generated by analysing the similarity results of the query sequence (~8 Kb ERV). Alphabets were incorporated into systems naming conventions in order to distinguish multiple hits on the same chromosome. Example PRT 5A system refers to a potential PRT system on chromosome 5 at a position A which is different from position B on the same chromosome. The two reference systems observed included PRT 5A and PRT 4B. Apart from possessing a high mismatch score and being highly specific in terms of amplifiable products, PRT 5A system also produced products with considerable size differences which could have served an added advantage. However, due to the presence of a sequence variant observed between the reverse primer binding site of the reference locus and the sequence derived from trace archives (Figure 24), further development of PRT 5A was discontinued. Since the same reverse primer was associated with PRT 10 K, PRT 1 E and PRT 4 B this G to A transition was also common for 10 K PRT and other systems which produced products differing in sizes from 5A and 10K PRT systems and therefore they were discarded as well.

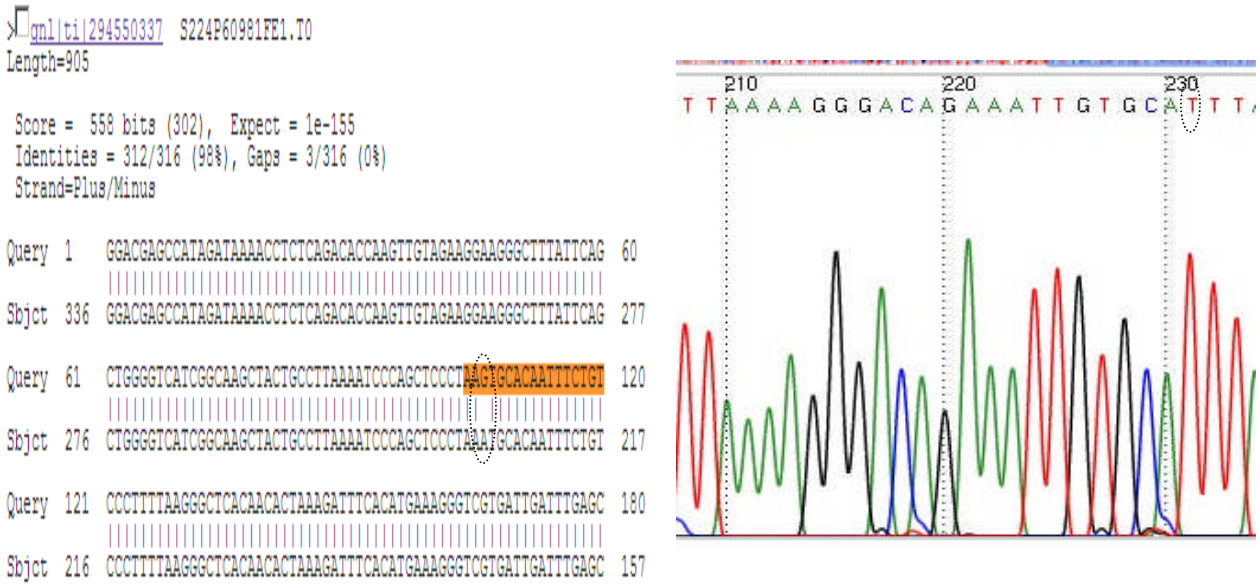
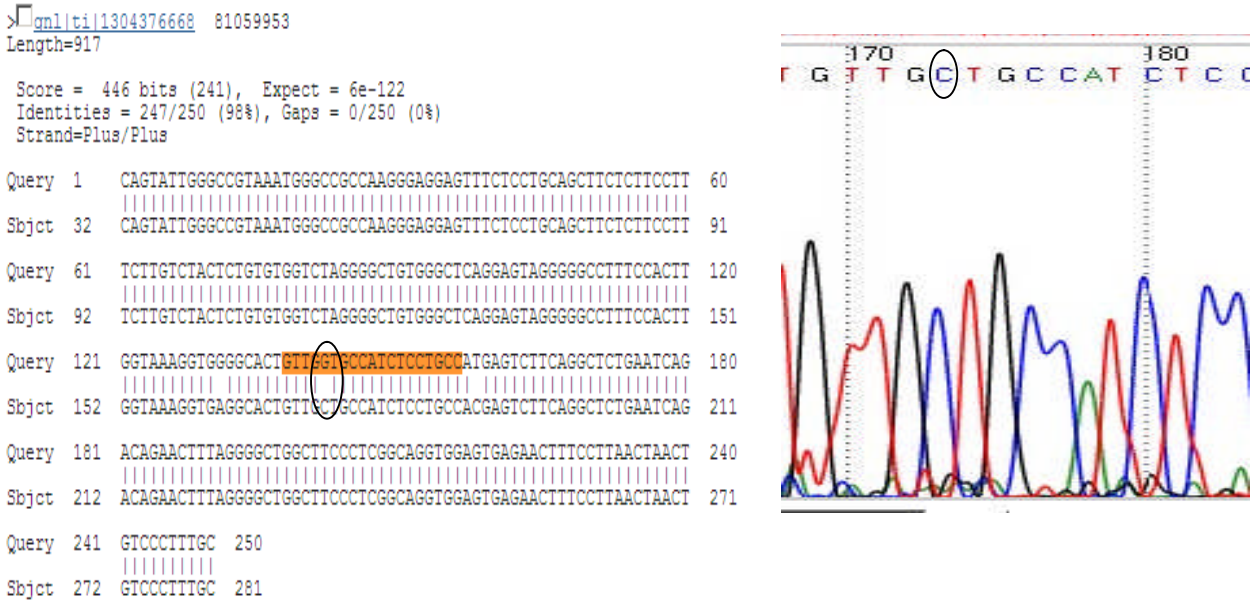


Figure 24. The sequence derived from trace archive for checking the presence of SNPs within the primer amplifying amplifying the reference sequence.

The sequence trace clearly depicts the confirmation of this SNP within the primer binding site.

Another two reference system identified was PRT 4B**.** A SNP was found to be localized within the reverse primer of this particular PRT system. The input query sequence showed high percentage sequence identity with the output trace archive sequence which added to the trustworthiness of the sequence and also to the existent transversion from G to C and hence this system was discarded as well.



>gnl|ti|1304376668 81059953
Length=917

Score = 446 bits (241), Expect = 6e-122
Identities = 247/250 (98%), Gaps = 0/250 (0%)
Strand=Plus/Plus

```
Query  1    CAGTATTGGGCCGTAAATGGGCCGCCAAGGGAGGAGTTTCTCCTGCAGCTTCTCTTCCTT  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  32   CAGTATTGGGCCGTAAATGGGCCGCCAAGGGAGGAGTTTCTCCTGCAGCTTCTCTTCCTT  91

Query  61   TCTTGTCTACTCTGTGTGGTCTAGGGGCTGTGGGCTCAGGAGTAGGGGGCCTTTCCACTT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  92   TCTTGTCTACTCTGTGTGGTCTAGGGGCTGTGGGCTCAGGAGTAGGGGGCCTTTCCACTT  151

Query  121  GGTAAAGGTGGGGCACTGTTCGTGCCATCTCCTGCCATGAGTCTTCAGGCTCTGAATCAG  180
            |||||||||| |||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct  152  GGTAAAGGTGAGGCACTGTTCCTGCCATCTCCTGCCACGAGTCTTCAGGCTCTGAATCAG  211

Query  181  ACAGAACTTTAGGGGCTGGCTTCCCTCGGCAGGTGGAGTGAGAACTTTCCTTAACTAACT  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  212  ACAGAACTTTAGGGGCTGGCTTCCCTCGGCAGGTGGAGTGAGAACTTTCCTTAACTAACT  271

Query  241  GTCCCTTTGC  250
            ||||||||||
Sbjct  272  GTCCCTTTGC  281
```


Figure 25. Another sequence variant present within the primer binding site of PRT 4B.

This PRT system was discarded as well due to the presence of SNP located within the forward primer.

3.1.2: Identification of three reference systems

Identification of three reference systems which includes PRTs which amplify three reference loci located at different genomic positions producing differently sized products along with the test loci. These systems included PRT 2A and PRT 4C. The three reference products amplified by PRT systems could not be distinguished by capillary electrophoresis because of the absence of size differences between individual test and reference loci amplified.

3.1.3: Identification of one reference PRT systems

Identification of one reference PRT systems which included PRT 4A, 12B and the selected PRT systems (12A and 1H PRT systems). A transition from C to T observed at quite high frequency which led to PRT 4A systems rejection.

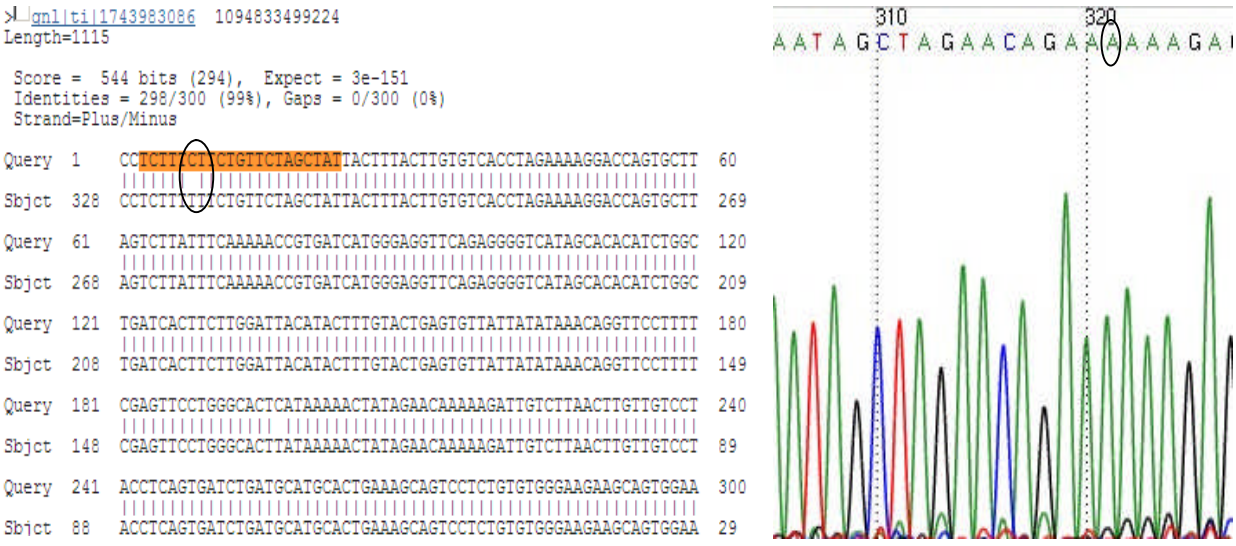


Figure 26. PRT 4A system which was rejected due to the presence of sequence variants within the primer binding sites.

PRT 12 B was also rejected as the size difference between the amplicons created by test and reference was small (147 bp versus 145 bp) and therefore this system was not given priority.

The finally selected PRT systems were one reference systems included PRT 12A and PRT 1H. No known polymorphisms were observed at the primer binding regions of these systems and enough size differences between the test and reference amplicons were predicted by PPPP. Moreover, the size differences between the test and reference loci were large enough

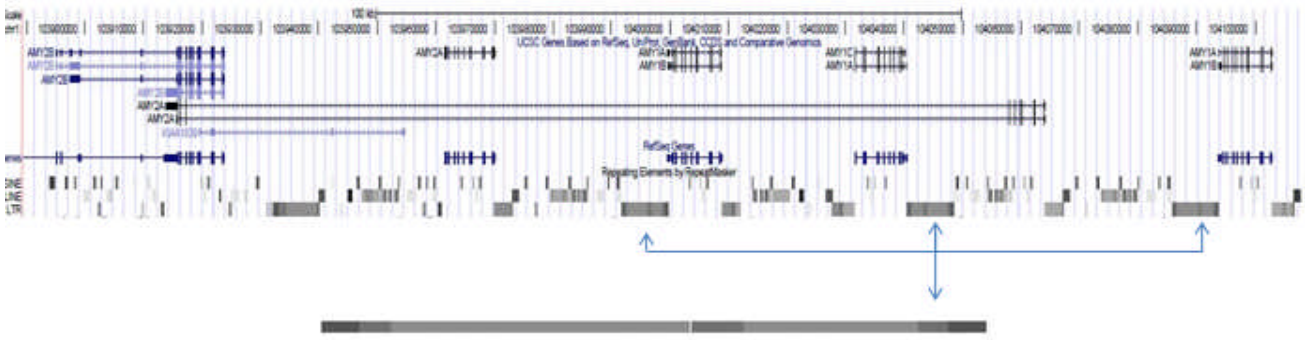
to be resolved over the capillary system. 1H PRT was observed to have size difference of 20 bp with test at 123 and reference at 142 and 12A PRT differed by 4bps in terms of test (244) and reference (249) amplicons. Hence they were selected for further assay development.

The melting temperatures of both forward and reverse primers were made to coincide by adjusting the primer sizes through UCSC's insilico PCR for rendering thermodynamic stability. As verification, the regions amplifiable by both 12 A PRT and 1H PRT were checked again by UCSC's Insilico PCR.

3.2: Genomic region: Location of 12A and 1H systems

Endogenous retroviral elements (responsible for salivary specific expression in case of salivary amylase) are scattered throughout the human genome and are the basis for the paralogous sequences selected for the construction of salivary amylase PRT. Therefore, these paralogous sequences also have a tendency to be dispersed throughout the genome exhibiting high degrees of sequence similarity with one another. Thus amplifying individual sequences without any contribution from other similar sequences can be a difficult task especially for a locus like salivary amylase because it exhibits considerable similarity (~96%) with pancreatic amylase gene copies. Amplifying just test and reference loci without any contribution from other similar sequence loci is an absolute essential characteristic feature of an accurate and reproducible PRT design.

In order to identify these sequences which exhibit high sequence similarity with the paralogous sequences used for salivary amylase PRT design, which can pose as a threat towards the accuracy of the system, mismatch scores for individual PRT systems were calculated based on the clustalW alignment.



IH PRT

CHR AMY1		1H Forward Primer 5'-3'	1H Reverse Primer 3'-5'	
(R)	141	CTTCTAAATCATAAGTTGATTT	AATAGAAAACAATATAAAACAAG	
6D	141	CTTCTAAATCATAAGTTGATTT	AATATGAAACAATATAAAACAAT	5
21A	141	CTTCTAAATCATAGGTTGATTT	AATATAAAACAATATAAAACAA-	6
7B	141	CTTCTAAATCATAAGTTGATTT	AATATGAAACAATATAAAACAAT	9
11A	141	CTTCTAAATCATAAGTTGATTT	AATATGAAACAATATAAAACAAT	5
YA	140	CTTCTAAATCATAAGTTGATTT	AATATGAAACAATATAAAACAAT	5
YC	141	CTTCTAAATCATAAGTTGATTT	AATATGAAACAATATAAAACAAT	5
11B	141	CTTCTAAATCATAAGTTGATTT	AATATGAAACAATATAAAACAAA	5
10C	141	CTTCTAAATCATAAGTTGATTT	AATATGAAACAATATAAAACAAA	5
IJ	141	CTTCTAAATCATAAGTGGATGT	AATAGAAAACAATGTAAAACAAG	6
1I	141	CTTCTAAATCATAAGTGGATGT	AATAGAAAACAATGTAAAACAAG	6
IK	141	CTTCTAAATCATAAGTGGATGT	AATAGAAAACAATGTAAAACAAG	6
19G	140	CTTCTAAATCATAAGTTGATTT	ACTATAAAACAATATAAAACAAT	7

12A PRT

CHR AMY1		12A Forward Primer 5'-3'	12A Reverse Primer 3'-5'	
(T)	244	ATCTAGTCCTTTTCTATCAATG	CTTGACAGATGTACTTATTTTCT	
5D	244	ATCTAGTCCTTTTCCATCAATG	CTTGACAGATGTACTTATTTTAT	2
1G	244	ATCTAGTCCTTTTTCATCAATG	CTTGACAGATGTACTTATTTTCT	4
1D	245	ATCTAGTCCTTTTCCATCAATG	CTTGACAGATATACTTATTTTTT	6
6A	247	ATCTAGTCCTTTTCCATCAATG	TTTGACAGATGTACTTATTTTCT	8
4B	243	ATCTAGTCCTTTTCCATCAATG	TTTGACAGATGTACTTATTTTCT	8
4E	237	ATCTAGTCCTTTTCCATCAATG	CTTGACAGATATACTTAATTTTT	8
5B	245	ATCTAGTCCTCTTCCATCAATG	TTTGACAGATATACTTATTTTTT	12

Figure 27. The genomic location of the salivary amylase gene cluster with 3 copies of salivary amylase associated with the retroviral element. The arrows depicted in this figure point to the retroviral element considered for the identification of test paralogous sequences. After deriving the potential primer binding sites through PPPP and selection of primers without SNPs, mismatch scores were calculated inorder to identify the potential loci which could be amplified through the same pair of primers producing test and reference. For each primer binding site, a score of 4 was given for each mismatch in the 3' bases. A score of 1 was given for mismatches in 5' and 2 for any mismatched bases in the middle of the primer. These scorings were essentially reflective of the ease with which mismatches could be tolerated by a primer pair.

3.3: Development of 12A PRT system

Appropriate annealing temperature was determined through gradient PCR and repeat testing was done for various standard cell and ECACC samples. It was initially visualized through agarose gel electrophoresis and then on the capillary with labelled primers which further verified the disappearance of extra loci with increase in annealing temperature.

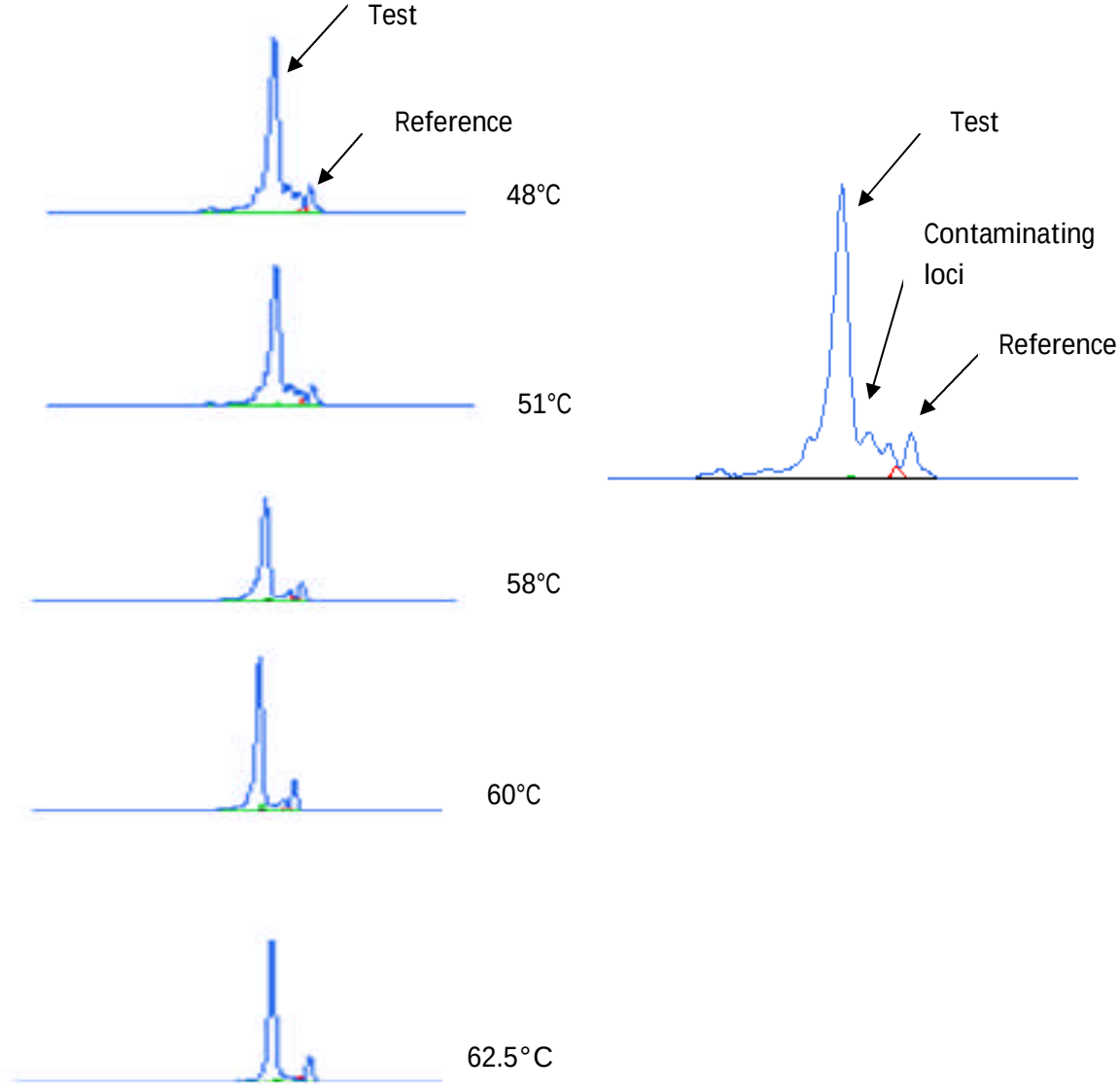


Figure 28. The discriminatory effect of increasing annealing temperature in PRT 12A. At higher annealing temperature only test and reference are produced.

3.3.1: Assessment of other contaminating loci by RFLP

The extent of amplification of additional loci coinciding in size with the test loci for PRT 12A was ascertained through restriction enzyme digestion. NEB cutter was utilized for identifying *Taq* I having a restriction site T/CGA at position 204 only for the test sequence,

not for reference and other amplifiable loci. The absence of any measurable product at the test locus after the restriction enzyme digestion confirms the lack of extra loci produced in the case of PRT 12 A (Figure 31).

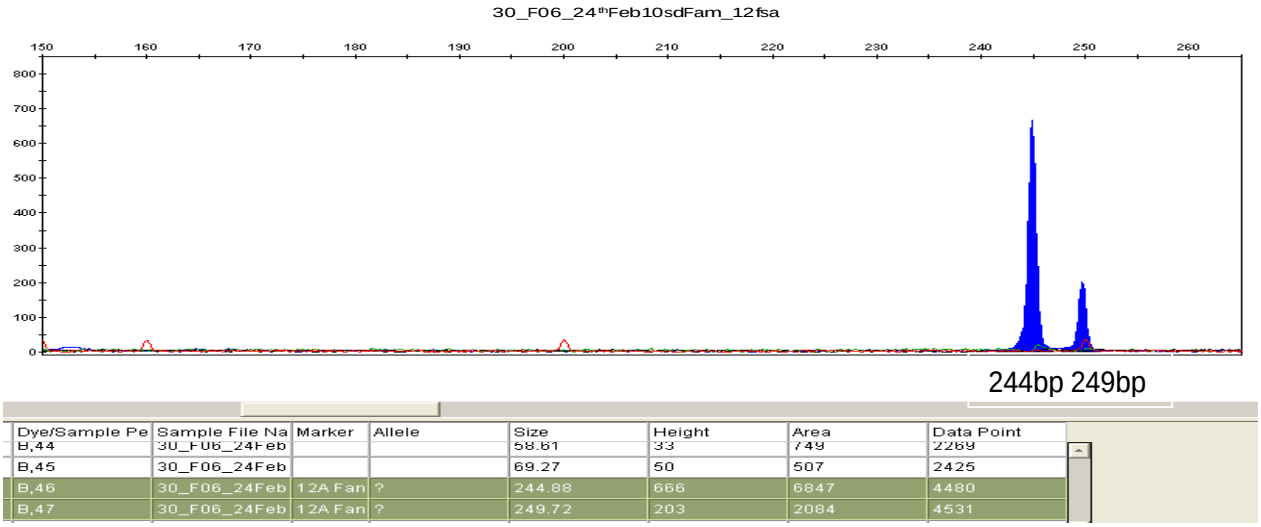


Figure 29.The amplified products of 12A PRT system with test at 244 and reference at 249.

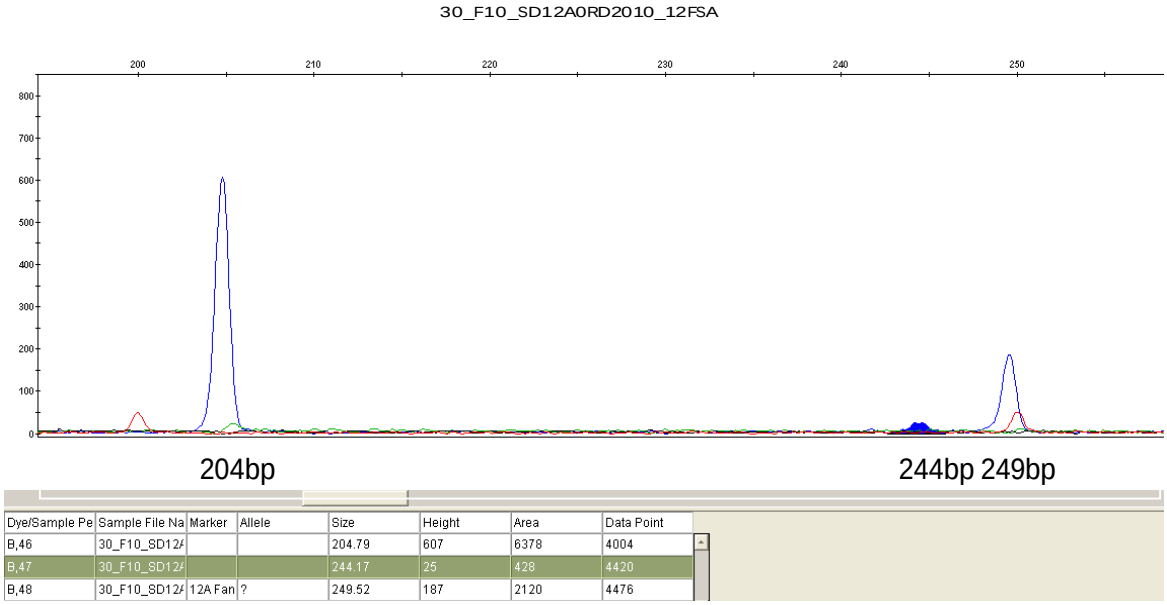


Figure 30. Taq I digestion of test sequence produces a product of size 204. Any Residual peak at 244bp represents contribution from additional loci.

3.4: Development of 1H PRT system

Appropriate annealing temperature was determined for 1 H PRT system. Test and reference bands were visualized on gel electrophoresis and later verified on the capillary.

3.4.1: PRT 1 H: Assessment of contribution towards the test locus by other loci

In order to determine whether amplification of additional loci contributed towards the reference product of 1 H PRT, restriction enzyme digestion by *Alu* I was carried out. *Alu* I possesses a blunt ended restriction site, AG/CT at position 26 of the reference product. The absence of cleavage at the reference locus of the 1H PRT system suggested either lack of digestion or amplification of extra loci.

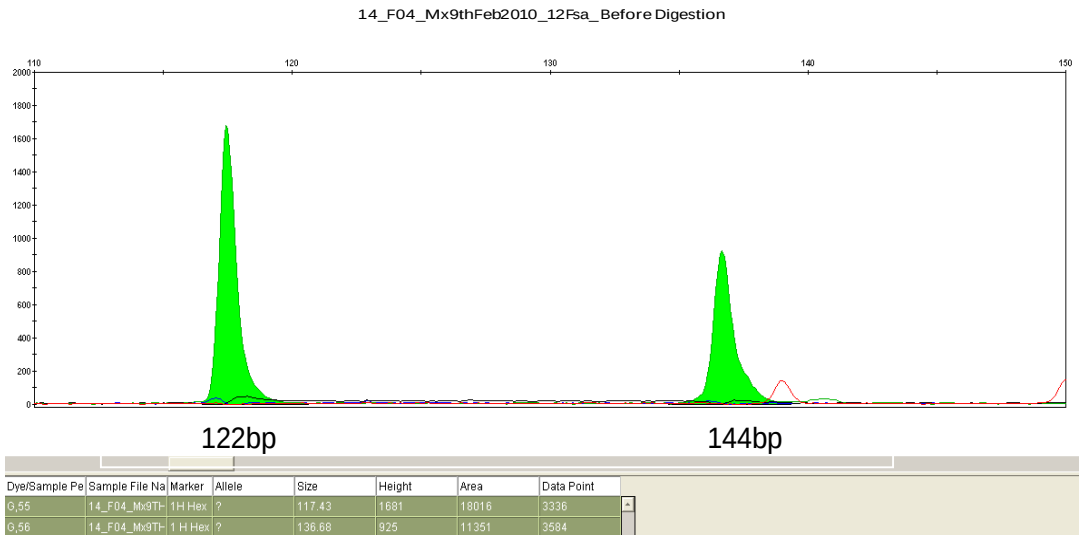


Figure 31. The amplified products of 1H PRT system. There was a discordance between bioinformatically determined sizes (Test 122, reference 141) and the sizes observed on the capillary which could be because of charge and mobility issues.

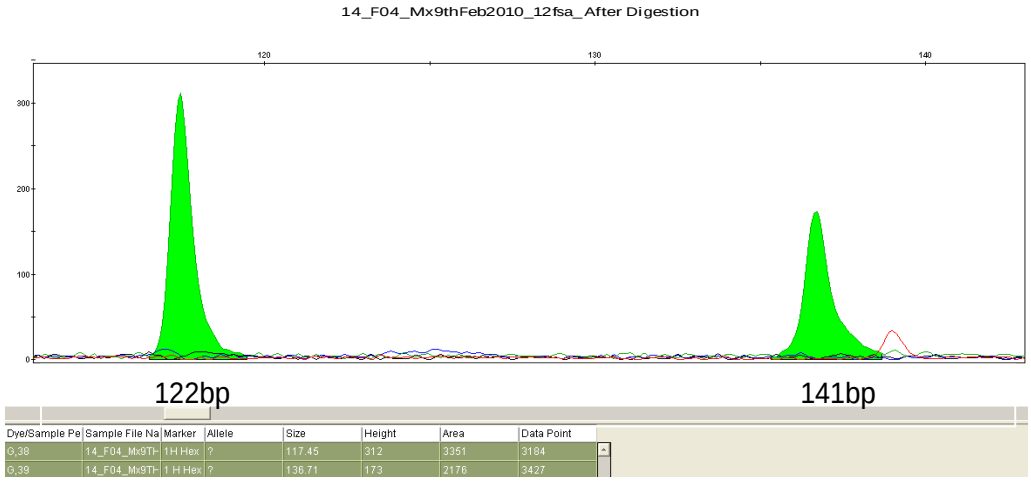


Figure 32. Alu I digested products of 1H PRT system indicating lack of digestion or amplification of extra loci. It was later confirmed that AluI digestion failed because of the formation of heteroduplex between test and reference sequences.

λ Hind III DNA ladder was digested with *Alu* I. It was ascertained that the enzyme *Alu* I was exhibiting optimal activity. Due to the consistency of ratios and associated copy number derived by 1H PRT system, it was very unlikely that alternative loci were getting amplified to such a great extent which was quantitatively hampering the reference product consequently leading to the undigested reference product. Therefore in order to confirm this hypothesis, bioinformatic sequence analysis was done.

3.4.1.1: Formation of Heteroduplex in 1H PRT

Bioinformatic sequence analysis of the test and reference loci revealed the potential creation of a heteroduplex between the test and reference loci which interfered with digestion capability of *Alu* I. Since this heretroduplex encompasses the restriction site of *Alu I*, it is very likely that the enzyme is not able to recognise its restriction site and cut the double stranded structure. Therefore, a procedure involving restriction enzyme digestion would not be helpful in identifying the extent of extra loci produced.

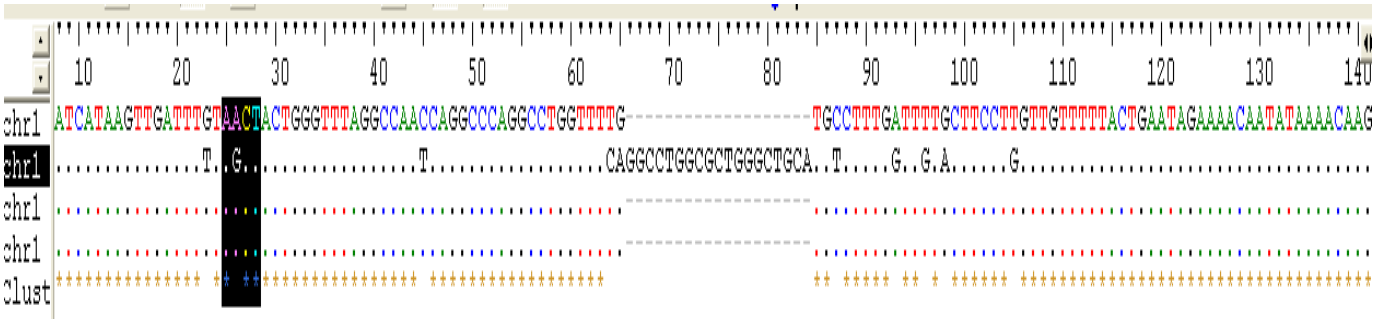


Figure 33. The first chromosome 1 represents the Test while the second one represents the Reference and the highlighted portion represents the restriction site for Alu I present at Reference (black square). There is considerable sequence similarity between test and reference at the start and ending of Test and reference products even though the middle appears quite different.

3.4.2: Assessment of other contaminating loci by direct sequencing

Since the sequenced PCR products of 1 H PRT would contain sequences of test, reference and other loci superimposed on one another, sequencing was carried out in order to identify

the potential contaminating loci. These sequences were later cross-checked with the sequence alignment containing test, reference and other loci.

Sequencing from the forward end revealed the presence of compression artefacts which produced certain odd spacing giving rise to incompatibility between the data produced by base calling software and the actual tracing visible. Therefore the traces were analysed again manually for identifying the correct order of individual nucleotides.

The sequence starting from GG at the position 39 of the alignment is common for all loci including test, reference and others. However, at position 46, a distinction in the alignment appears between other loci and the test- reference sequences which can be seen even within the sequenced samples.

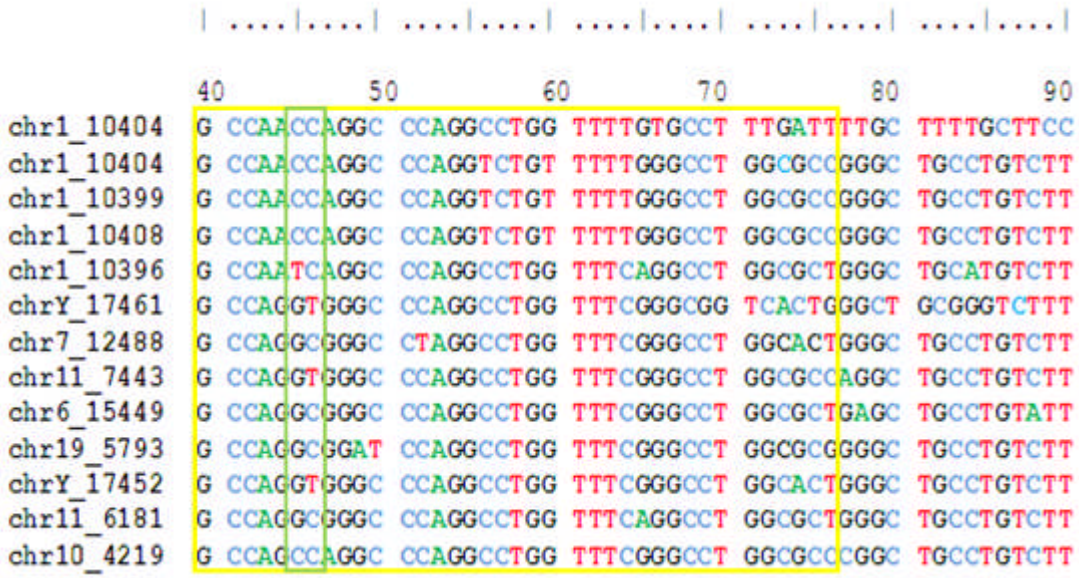


Figure 34. Sequence alignment containing target regions of test, reference and contaminating loci also observed in the sequence trace of multiple samples.

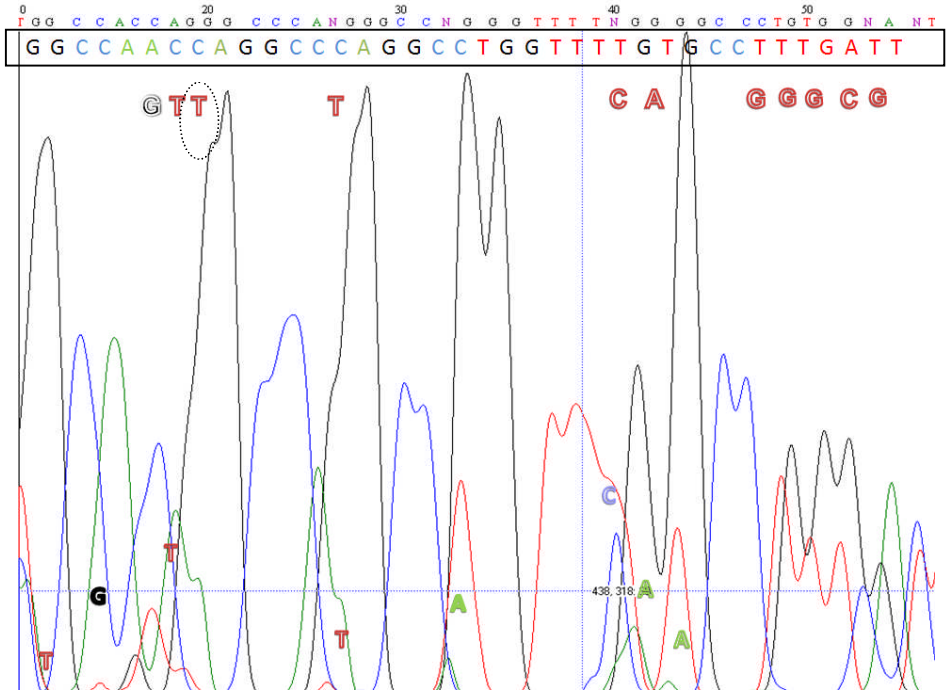


Figure 35. The sequence trace of the products amplified by 1H PRT which contains test, reference and contaminating loci superimposed on each other. Position 46 acts as a marker for identification contaminating loci.

The sequence at this location can act as a marker position for identifying the extra loci getting amplified. The marked T represents the position 46 according to the alignment and highlighted in green, present only in case of amplification of chromosome Y or chromosome 11.

Since chromosome Y is absent in females, further analysis of this region proved that both were amplified but the extent of amplification was less as demonstrated by the sequenced trace for males and females.

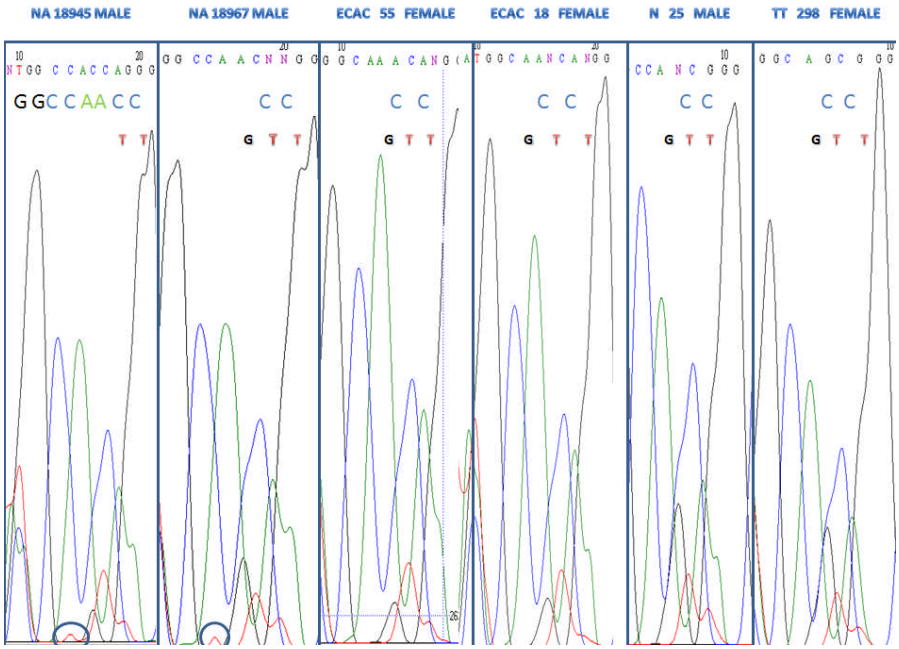


Figure 36 Analysis of the 46 marker position separately in males and females to ascertain the amplification of contaminants, chromosome Y or chromosome 11.

Intensity of the nucleotide T acts as a marker in females specially for identifying the extra loci being amplified. According to the alignment nucleotide T is present only on chromosome Y and chromosome 11 but absent from all other loci and since chromosome Y is absent in females, the presence of nucleotide T in sequencing analysis in females suggests dual amplification by chromosome 11 and chromosome Y in case of males which confirms that these two loci are produced but to a lesser extent.

3.5: TG Microsatellite (104051096-104051137, chromosome 1, March 2006 Assembly)

Microsatellite identification and further development was carried out for corroborating the copy numbers already established by the individual PRT systems. However, it needs further investigation with regard to the profile exhibited by it.

3.5.1: Genomic Region

A TG microsatellite was found on chromosome 1 through UCSC genome browser. UCSC's BLAT was also utilized to verify its exclusive association with AMY 1 only. The genomic

region encompassed by this particular microsatellite is located in the close vicinity of the retroviral element associated with one of the salivary amylase genomic copies.

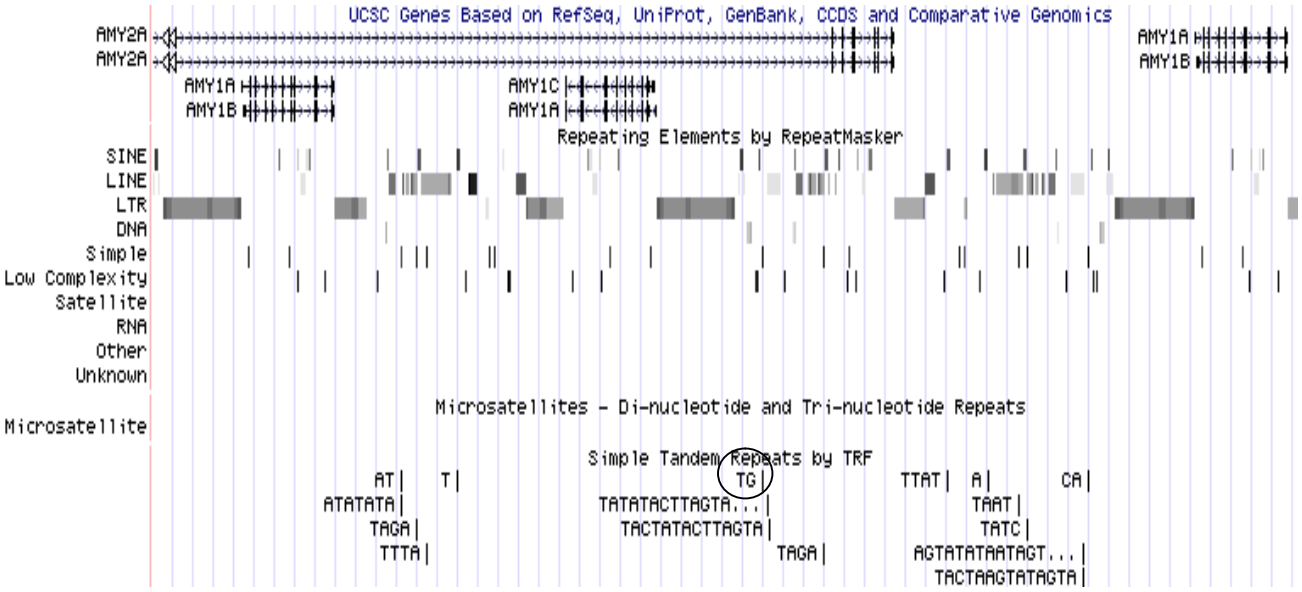
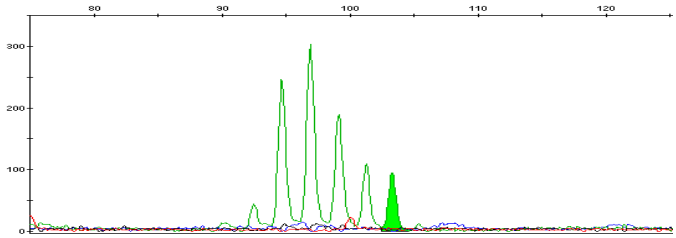


Figure 37. The genomic location of the TG microsatellite.

3.5.2: Development of TG microsatellite assay

Primers for microsatellite analysis were designed through Primer3 software. The size of the microsatellite amplicon was chosen to avoid coinciding with 1H PRT and 12 A PRT amplicons to allow successful multiplexing.

Additionally, the primer regions were checked for polymorphisms through trace archives and polymorphism in the number of dinucleotide repeats was noted. Also, for ensuring exclusive amplification of salivary amylase, enough mismatches were made to coincide within the primer binding sites for exclusive amplification of the TG microsatellite within the AMY1 locus. These primers were later checked on agarose and then on capillary electrophoresis but the occurrence of stutter bands made certain traces difficult to interpret.



Dye/Sample Pe	Sample File Na	Marker	Allele	Size	Height	Area
G,126	03_C01_8thMa			89.99	14	210
G,127	03_C01_8thMa			92.42	44	452
G,128	03_C01_8thMa			93.34	13	108
G,129	03_C01_8thMa			94.64	246	2440
G,130	03_C01_8thMa			96.88	303	3046
G,131	03_C01_8thMa			99.14	189	1983
G,132	03_C01_8thMa			101.3	109	1147
G,133	03_C01_8thMa			103.25	95	984

Figure 38 Amplified products of the TG microsatellite in a standard cell line sample (AF103). Although the genome browser predicted a biallelic profile according to the primers selected by Primer3, this particular sample exhibited a multiallelic profile. According to the PRT analysis, this particular sample has 3 copies which could mean that either the observed multiple peaks were due to PCR slippage or this sample genuinely has a multiallelic profile.

Although the genome browser predicted a biallelic profile for the amplified microsatellite product, the occurrence of multiple peaks particularly 1-5 repeat units smaller than the main allele suggested either a multiallelic profile or excessive slippage. Due to the presence of multiple peaks on the capillary, the original two alleles as predicted by the genome browser could not be analysed.

Since the main role of microsatellite involves confirmation of copy numbers as determined by the PRT assays and also due to the uncertainties associated with the size of the duplication unit affecting the microsatellite's profile, preference was given to first establishing accurate and reproducible PRT systems.

3.6: Further development of PRT Assays

The PRT assays were first tested on standard cell line, ECACC and Japanese samples. Repeat testing confirmed consistency of ratios for peak height and peak area between experiments done on different occasions.

3.6.1: Calibration of Reference Samples

Japanese samples were used as reference samples because published copy number data was available for these samples and they were accessible in the laboratory. They were calibrated

with the copy number data ascertained by Perry (Perry et al, 2007) for the same set of samples mostly through real time PCR (one sample NA 18972 exhibiting 14 copies of salivary amylase ascertained through Fibre FISH). Initially all Japanese samples were typed by both PRT systems and consistency between individual sample ratios on different occasions was noted. Individual samples representing each copy number as typed by P.J. Perry were plotted against the ratios determined by both PRTs for the same set of samples. Since there was a strong relationship between copy number determined by P.J. Perry and ratios determined by PRTs for the selected set of samples even with repeat testing and this relationship was later utilized for the assessment of copy numbers based on their ratios for other samples.

3.6.1.1: Selection of Reference Samples

The Japanese reference samples representing high copy number classes (from 4 copies to 14 copies) were selected on the basis of copy number agreement between both the PRT systems and P.J. Perry's typed samples. The figures mentioned below represent a linear relationship between ratios and copy number which is seen to be consistent with repeat testing. These experiments done on different dates confirm the consistency of the relationship between ratios and copy number for both 1 H (HEX) and 12 A (FAM) PRT systems in spite of some differences in the relationship expressed by peak height or area.

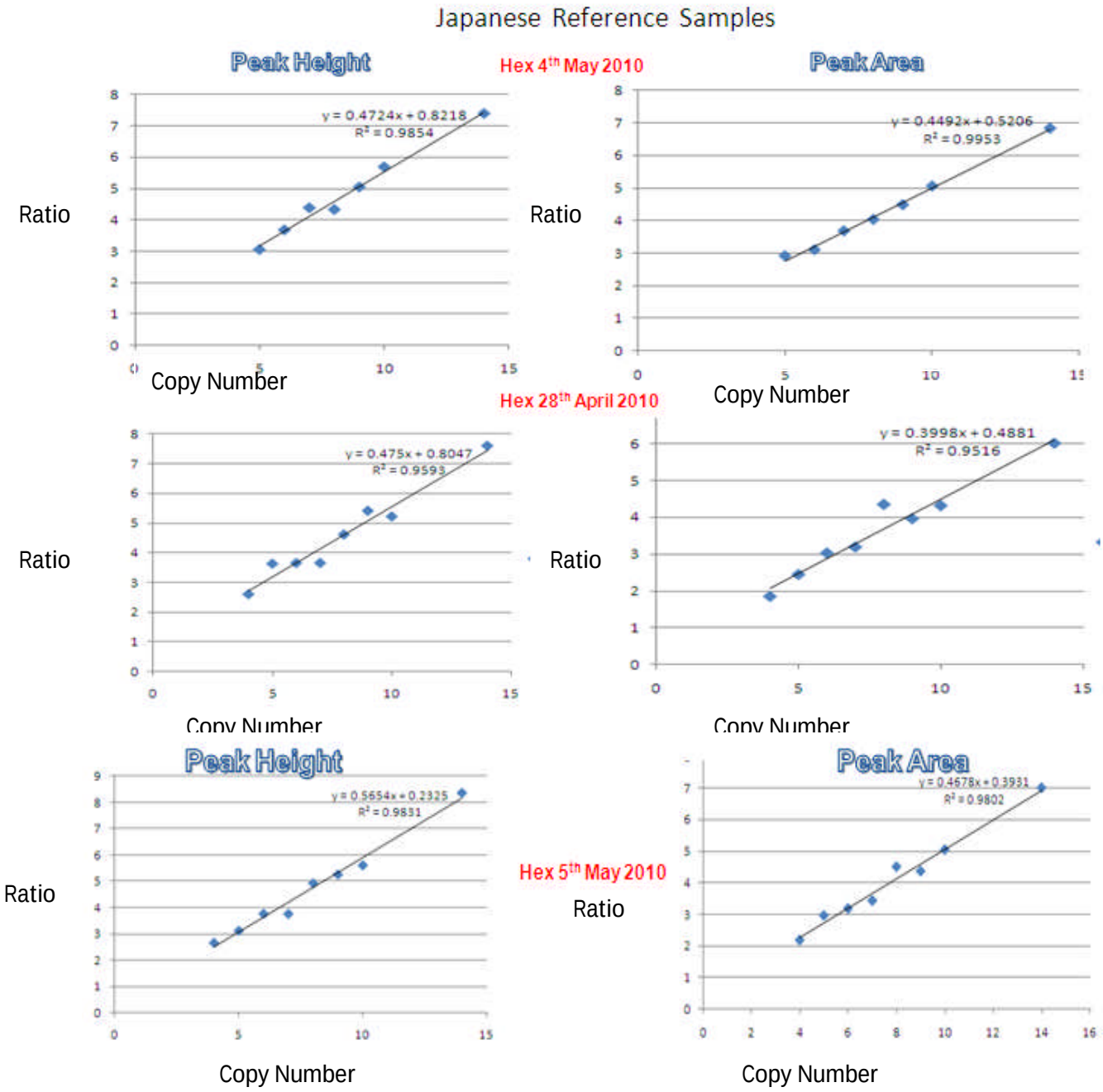


Figure 39. Linear plots of copy number with respect to ratios for 1H PRT system based on peak height and peak area. Repeated high correlation between copy number and ratios substantiates the accuracy and precision of 1H PRT system.

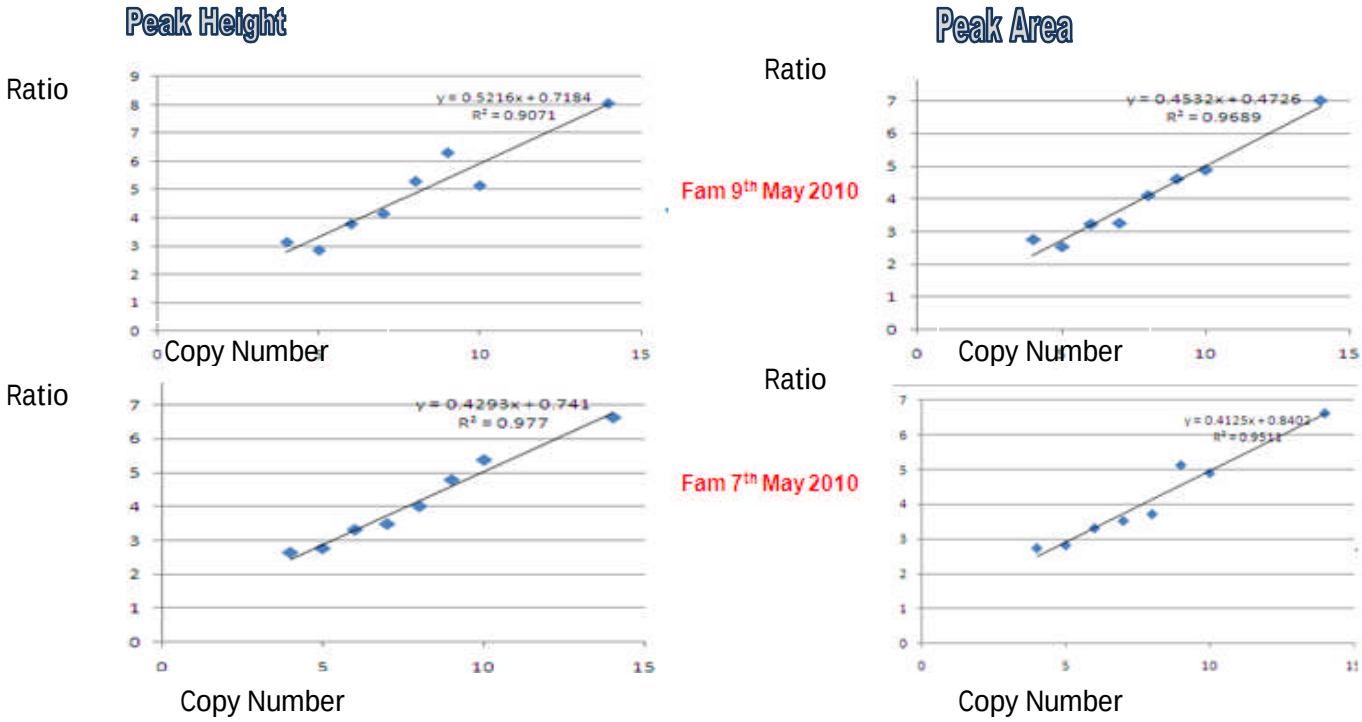


Figure 40. The Relationship between copy numbers and ratios for 12A PRT system. The strong correlation observed repeatedly is indicative of the accuracy of this system.

3.6.2: Agreement between both 12A and 1H PRT systems

The reference repeat units for these PRT systems are located on two different chromosomes therefore, evaluation by these two systems can be considered as measurements from two individual sources. Although numerous combinations with respect to peak height and peak area for both the PRT systems are possible, agreement between both the systems for all possible combinations was noted along with a variation in agreement particularly with regard to height and area.

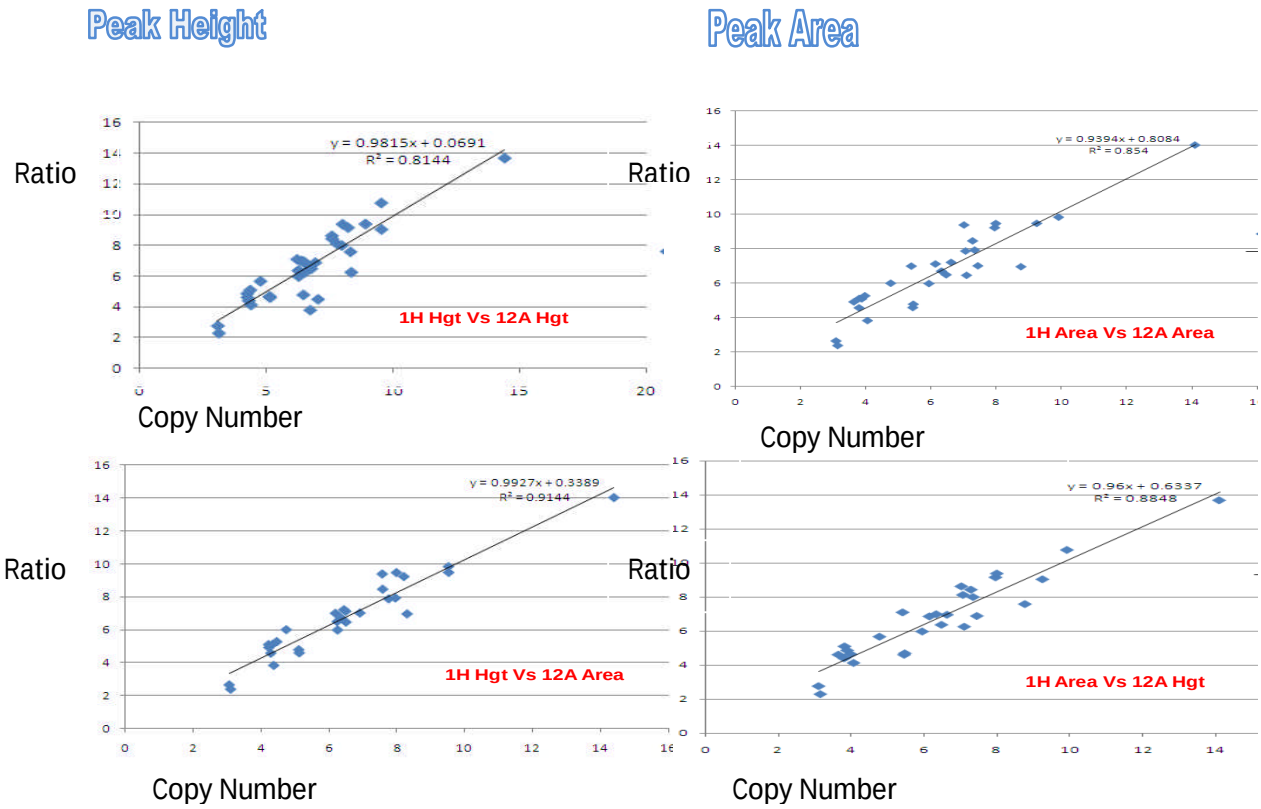


Figure 41. The copy numbers derived from 1H and 12A according to the peak height and area. The strong correlation derived from two individual measurement systems was observed to be consistent for all possible combinations.

3.7: Comparison of peak height and peak area

In order to identify a more accurate measure between peak height and peak area, regression plots involving test height and test area; reference height and reference area were produced. The idea behind it was that any "normal peak" would always be associated with a well proportioned characteristic shape dependant on the electrophoretic run of the capillary and in such a situation height and area of individual peaks would always exhibit a linear relationship with one another. Any kind of deviation from the line denoting the relationship would act as a marker to identify peaks with distorted shapes which is often associated with excessive discordance between height and area ratios.

As seen in below, reference peaks were observed to follow a linear trend with no outliers but the test peaks for both the PRT systems encompassed outliers. Further investigation and examination confirmed that the samples exhibiting excessive discordance between height and

area ratios were also indicative of distorted shapes. On an average about 6% of the samples showed deviation for 12 A PRT and about 3 % showed deviation from the regression plot for

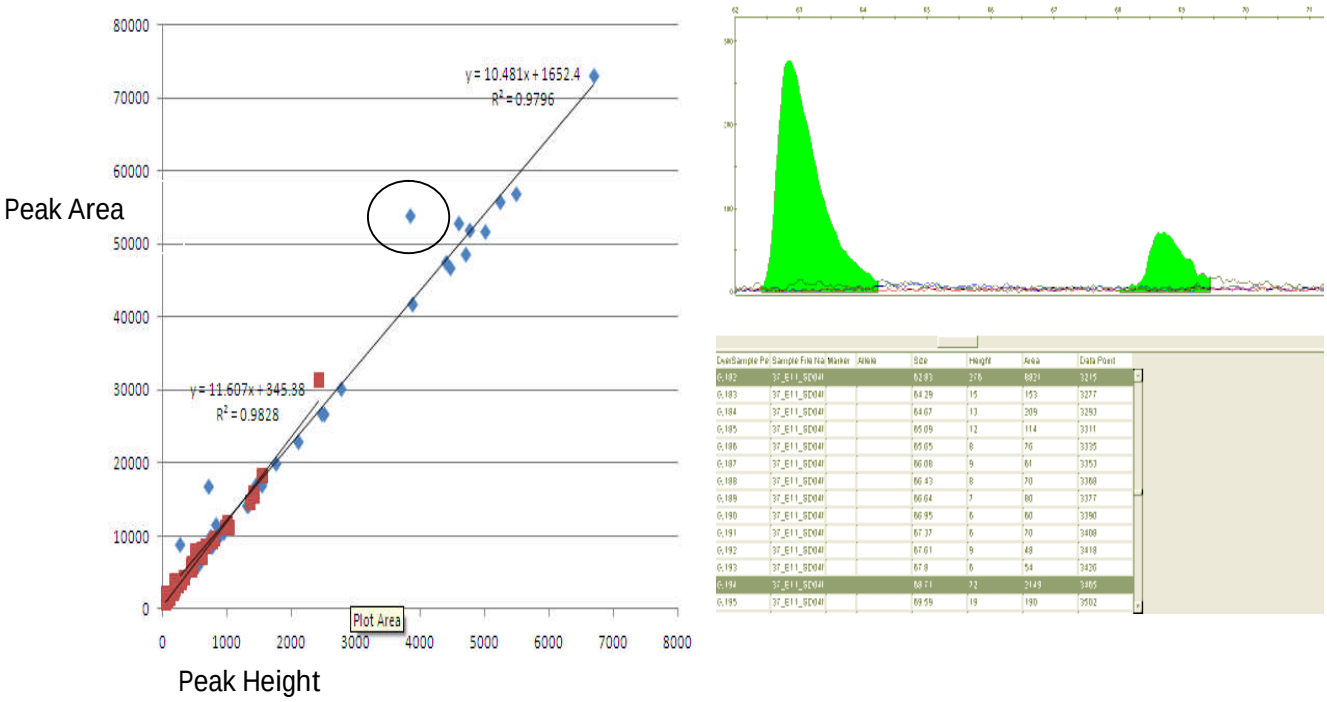


Figure 42. Linear scatter plot between test peak height and test peak area. (seen in blue, the red dots denote reference peak height and test peak area for 1H PRT system) The marked sample showed deviation from the rest of the data points due to shape distortion as highlighted by the diagram on the right.

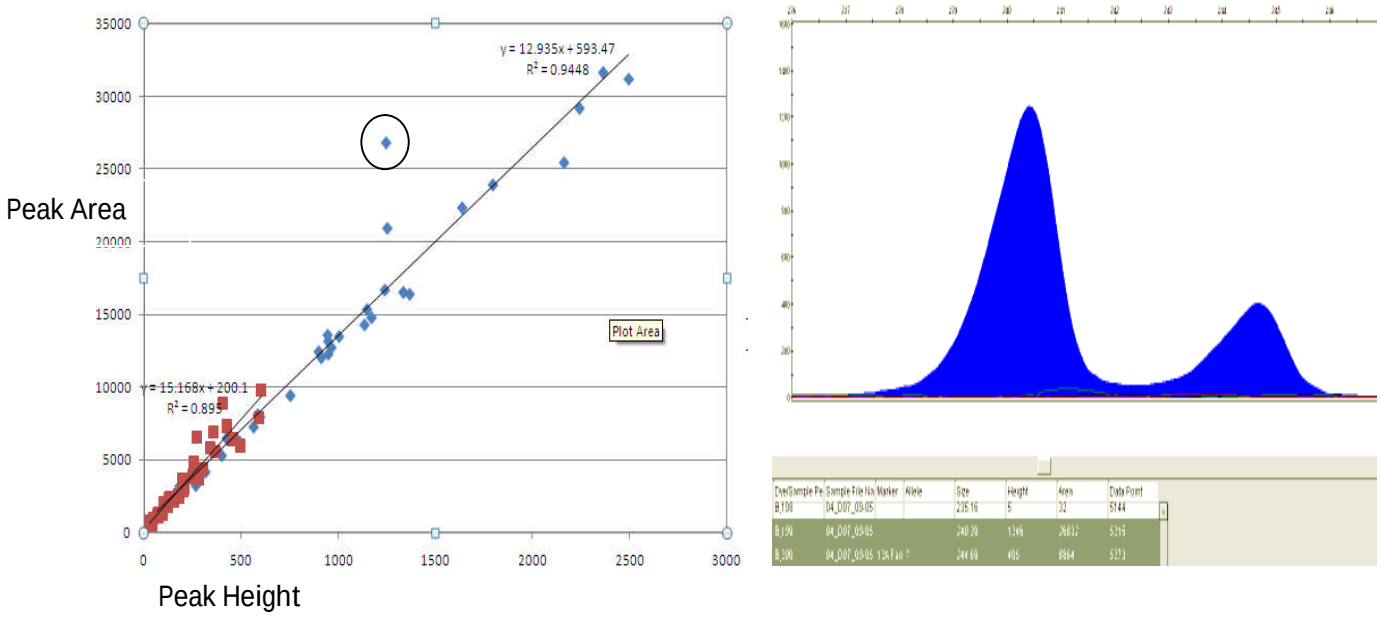


Figure 43. Linear scatter plot between test peak height and test peak area as seen in blue, the red dots denote reference peak height and test peak area for 12A PRT. The

outlier samples when seen individually denoted admixture between test and reference peaks (as seen in the diagram on the right) thus giving an incorrect ratio and corresponding copy number.

For assessing the extent of deviation eventually leading to discordance between area and height, quality control measures were established defining a set of parameters for differentiating between acceptable data and unacceptable data. Therefore, frequency distribution curves involving the ratios representing test peak area divided by test peak height were plotted against frequency thereby assisting in the identification of outliers within the dataset for both PRT systems.

3.8: Distribution of data points

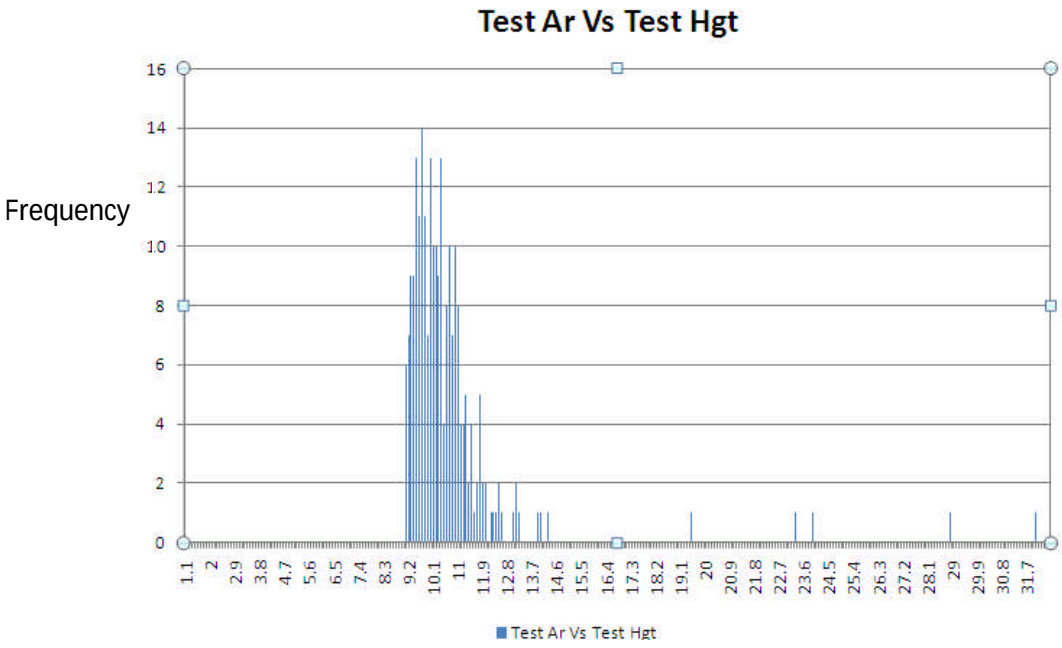


Figure 44. Frequency distribution of ratios of test peak against test height for 1H PRT thus identifying the outliers within the dataset.

Since the distribution was a unimodal, it was observed that in spite of being skewed the distribution was a good fit for normal distribution curves.

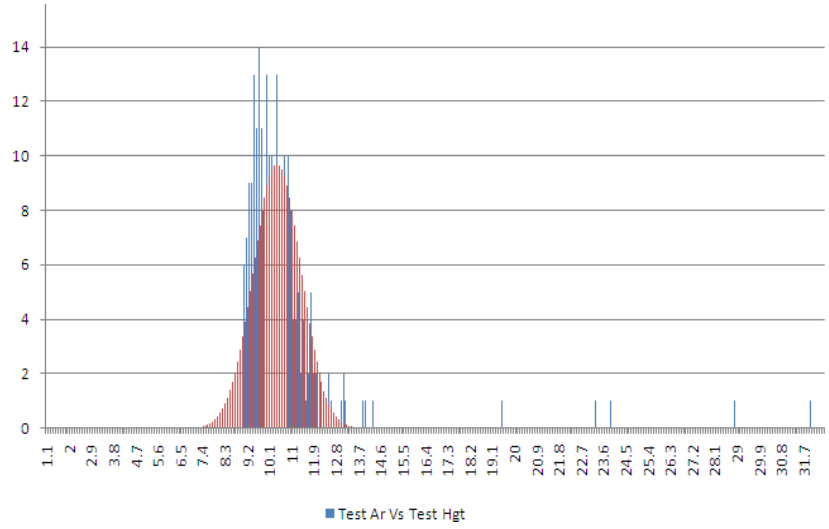


Figure 45. Simulation of the data set for 1H PRT in order to ascertain the potential outliers and establish a quality control.

3.9: Distribution of data points for 12A PRT

As the data points for 12 A PRT reflect a bimodal distribution which could be because of independent unimodal forces acting on it, therefore characterising the extent of variation for this particular PRT needed further analysis.

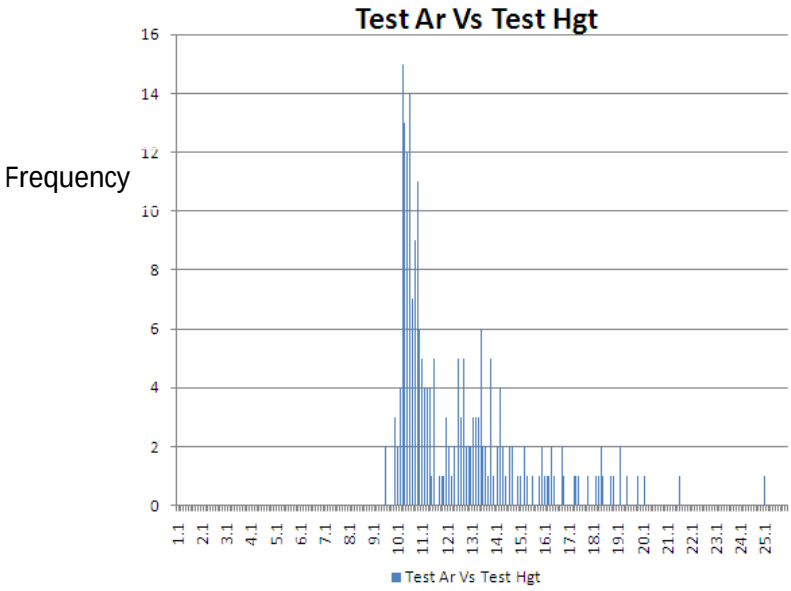
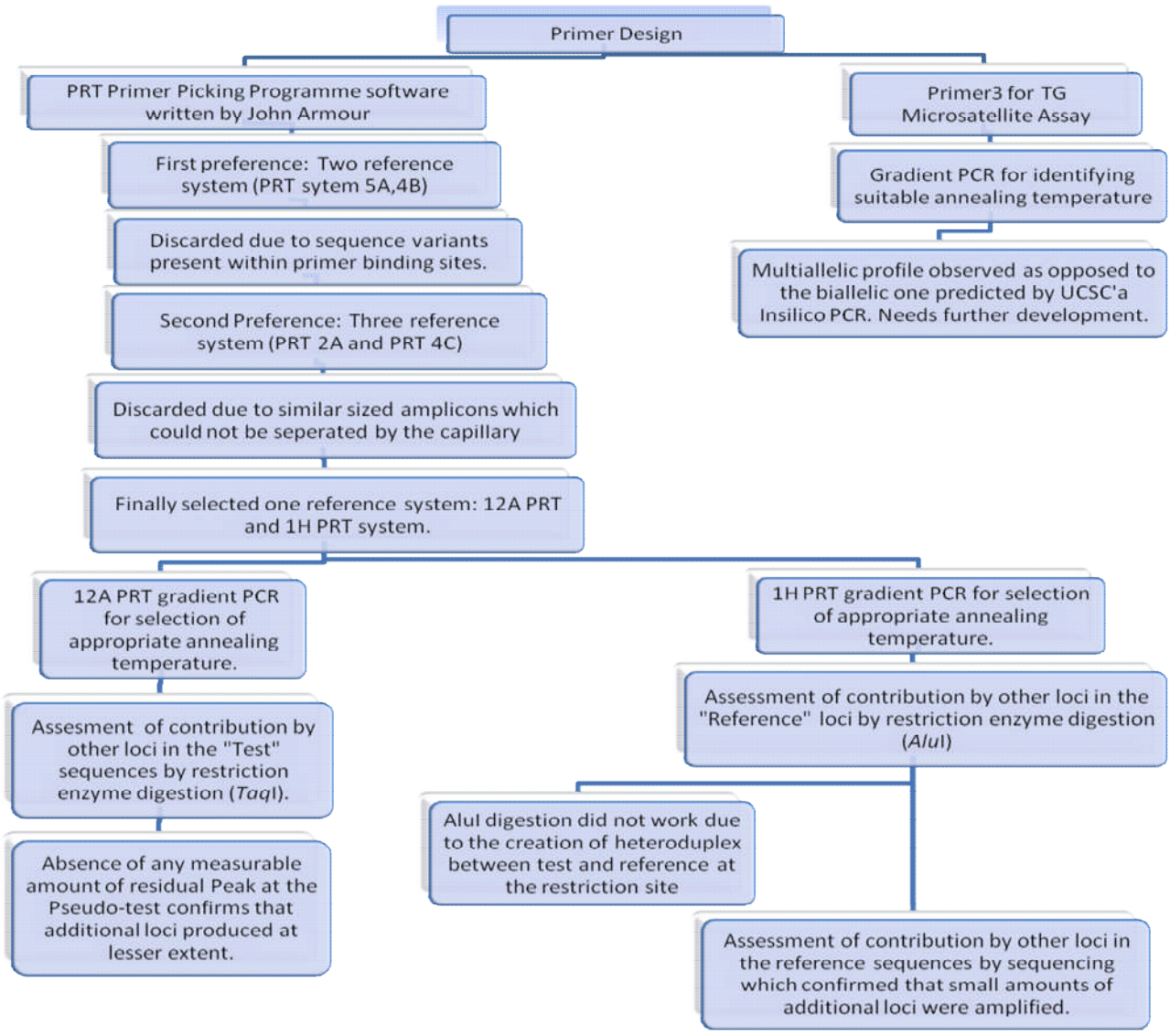
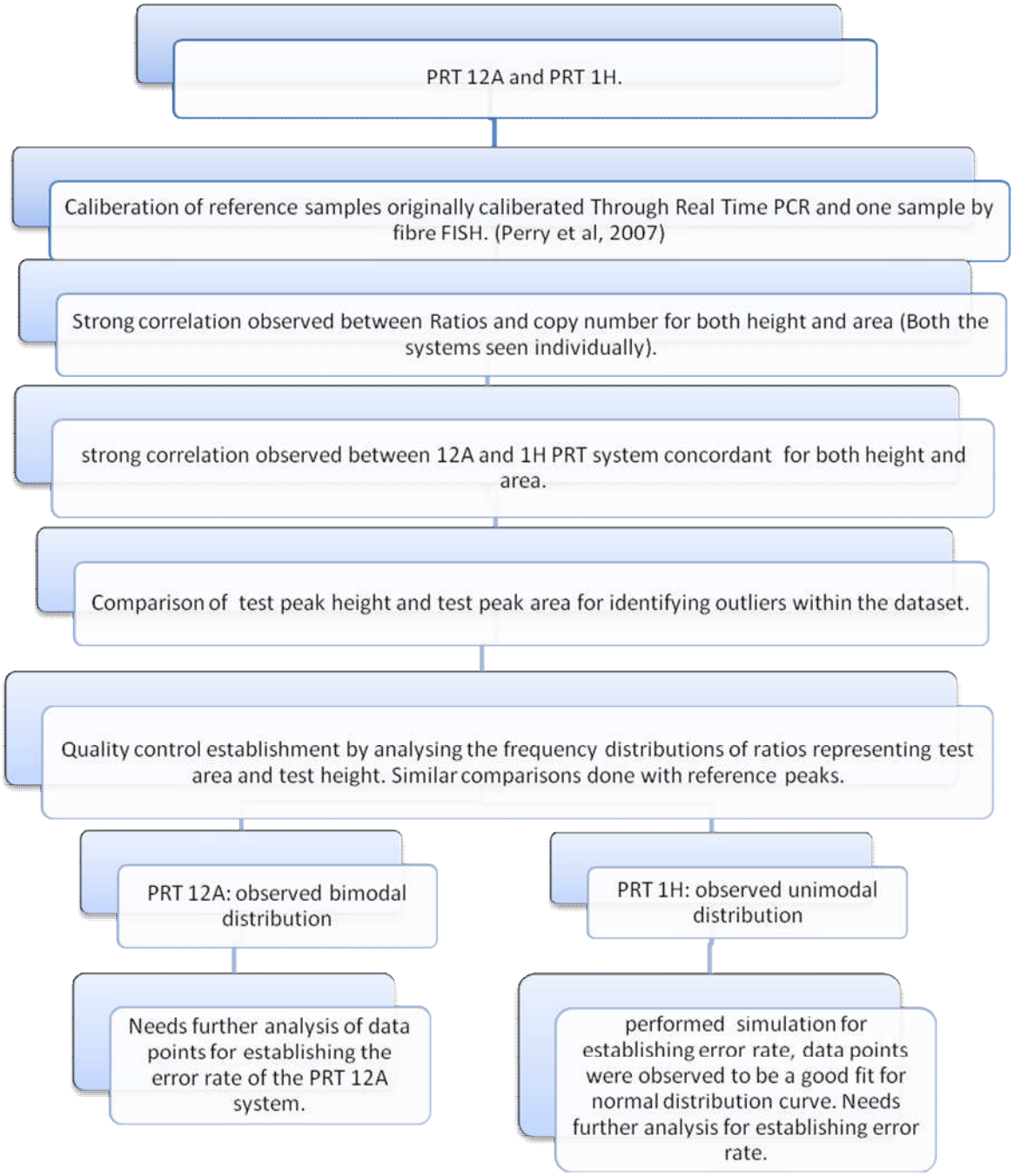


Figure 46. The frequency distribution of ratios of test peak against test height for 12A PRT thus identifying the outliers within the dataset.

3.10: Flowchart summary:




PRT 12A and PRT 1H.
Caliberation of reference samples originally caliberated Through Real Time PCR and one sample by fibre FISH. (Perry et al, 2007)
Strong correlation observed between Ratios and copy number for both height and area (Both the systems seen individually).
strong correlation observed between 12A and 1H PRT system concordant for both height and area.
Comparison of test peak height and test peak area for identifying outliers within the dataset.
Quality control establishment by analysing the frequency distributions of ratios representing test area and test height. Similar comparisons done with reference peaks.
PRT 12A: observed bimodal distribution
PRT 1H: observed unimodal distribution
Needs further analysis of data points for establishing the error rate of the PRT 12A system.
performed simulation for establishing error rate, data points were observed to be a good fit for normal distribution curve. Needs further analysis for establishing error rate.

CHAPTER 4: Discussion

The project initially concentrated on developing accurate and high throughput salivary amylase (AMY1) copy number measurement systems based on paralogous ratios derived from the amylase test locus and a reference locus exhibiting high sequence identity. Since AMY1 exhibits high copy number in the range of 2-14 copies per diploid genome achieving accuracy with reproducibility is difficult and inaccurate measurement systems may conceal the true extent of functional correlation. Amplification by means of one primer pair ensured the removal of variation caused by relative amplification efficiency of test/reference amplicons which aided in the development of accurate copy number measurement systems. Owing to the complex structural nature of copy number variations and consequent problems pertaining to the reproducibility and accuracy of the methods currently available to quantify them, the AMY1 PRT systems developed in this study combine accuracy with high throughput. The exclusive association of the provirus with the salivary amylase (AMY1) copies ensured that only salivary amylase copies were amplified and not the pancreatic amylase (AMY2) copies despite high sequence identity between them. As the proviral insertions are scattered throughout the human genome, one of the technical challenges affecting the reproducibility and accuracy of the assay was the non amplification of these alternative loci which was accomplished by the precise placement of primers within the target regions. Discrimination by means of increasing annealing temperature ensured exclusive amplification of only test and reference loci. For the 12A PRT system, restriction endonuclease digest experiments were utilized for assessing the potential contribution of amplification of other alternative loci. As the restriction enzyme *Taq I* had restriction sites for recognising only the test sequence and not the reference and other alternative loci, a clear distinction between test and alternative loci was made. The absence of any measurable product at the “pseudo-test” locus further substantiated the evidence against the amplification from other loci apart from test and reference. As for the 1H PRT system, experiments involving restriction enzymes could not yield conclusive results because of heteroduplex formation within the restriction site and consequent recognition and cleavage failure. Sequencing of 1H PRT’s PCR products revealed traces of all amplified products superimposed on one another and analysis of these products revealed the absence of any measurable amount of products produced by alternative loci amplification. In addition to the PRT systems, allelic ratios from a dinucleotide microsatellite present in the copy number

variable region were also initially developed so as to verify the integer copy number ascertained by individual PRT systems. However, due to the uncertainties with respect to its profile, further investigation is required.

In all, 2 PRT systems and 1 microsatellite system were developed to measure the copy number of salivary amylase genes in 48 Japanese Hap Map samples exhibiting high copy number range (4-14 copies) already typed by Perry (Perry et al, 2007) by real time PCR. These were divided into batches of 16 samples for convenience. Ratios of the initial 16 samples tested by both PRTs repeatedly displayed a strong relationship with the copy numbers for the same set of samples as deduced by Perry (Perry et al, 2007) in particular one sample (NA18972 with 14 copies) measured by fibre FISH was observed to have 14 copies when measured by both PRT systems highlighting the accuracy of both the systems. Reference samples specifying individual classes of integer copy number from 4 copies to 14 copies exhibited strong relationship with the copy numbers determined by Perry (Perry et al, 2007) and were selected for calibration for the rest of the samples.

The occurrence of a strong correlation observed between ratios of selected reference samples and Perry's corresponding samples, even after repeated testing reflects the existence of a linear relationship between ratios and copy number concordant for both PRT systems
A slight variation in the ratios determined by peak height and peak area was also observed for certain reference samples. The discordance between height and area ratios might be because of capillary artefacts which could be eliminated by rerunning the respective samples or because of amplification conditions eventually producing distorted shapes of peaks.
In order to identify more accurate basis for measurement from peak height or area, regression analysis was performed involving ratios derived by considering height and individual copy number. The same was done for peak area and it was concluded that area ratios for 1H PRT system were more reflective of their copy number classes and for 12A PRT system, peak height ratios was a more accurate measure. For 12A PRT, regression between test area and test height substantiated the accuracy of the height ratios by identifying samples with characteristic shapes highlighting the admixture between test and reference peaks, thus contributing to more area than accounted for. It was also suggestive of the extent of deviation causing discordance between area and height.

Even though each run of a capillary is empirically different however the combined data reflected the extent of deviation thus identifying certain outliers within the data set for both the PRT systems. In future, it would be beneficial to establish quality controls based on this dataset which would add to the efficacy of the PRT assays. For quantifying the amount of alternative loci produced by individual PRT systems, a primer extension assay (SNaP shot) (Brown et al, 2001) could also be utilized. Other parameters which could be included to judge the performance of both the assays include defining error rates, and "specific deviation", which elucidates the copy number of a particular sample by defining a normalised average deviation of an unrounded value with respect to its expected integer copy number. However, since these parameters could only be used for samples with known copy number, PRT-determined copy number could be verified by other techniques which would add validity to the results. Other techniques such as Fibre FISH or modifications of PRT could be used. A major disadvantage of PRT technique is its reliance on paralogous sequences scattered within the genome which are used as a reference for quantifying the copy number variable loci, however such sequences might or might not be present within the genome. It would perhaps be useful to develop a technique which could use other related sequences to quantify the test sequences such that even specific loci which do not share paralogous within the genome could also be quantified accurately. As opposed to other techniques, this would prove out to be high throuput and flexible.

As the microsatellite assay developed is less informative, other PRT systems having a reference locus on other chromosomes could be developed which would provide independent measures of the same test locus. Multiplexing and using different fluorescent labels would further reduce the cost and add to the convenience of the systems. Though inevitably there will be differences in copy number assessment by individual PRT systems which could either be because of random error or because of the fact that these PRT systems genuinely measure different copies. This could be discerned with the help of computer programmes distinguishing different systems based on weighted preference. However, discordance could also be because of the presence of polymorphisms within the primer binding sites leading to non amplification of specific copies of test (leading to depletion of ratios and prediction of a lower copy number) or reference DNA (leading to ratios being inflated and prediction of a higher copy number) and inclusion of a partial repeat in one system but not in the other might also lead to discrepancies.

Due to the presence of SNPs discovered within the primer binding sites amplifying the two reference loci and one test locus, two reference PRT systems were discarded. However using custom made mixed nucleotides positioned on the SNP would theoretically amplify the target region without dropping individual copies. Another potential line of investigation particularly for assessing higher copy number samples includes creating PRT systems with the single reference locus at a fixed higher copy so that the relative difference in signal strength between test and reference is reduced and regression can then be performed to decipher integer copy number of the samples.

As opposed to the argument regarding the presence of non integer copy number for amylase, the PRT experiments concluded the occurrence of integer copy numbers highlighted by the presence of clusters derived by independent measurements of the test locus. Moreover, the entire concept of adaptive evolution whereby positive selection favoured the divergence and expansion of the salivary amylase multigene family within humans goes against mosaicism. Using segregation analysis to ascertain the haploid copy number would not perhaps be as useful particularly because of the high amount of permutations and combinations possible for different haplotypes of amylase (diploid copy number as high as 14).

As salivary amylase copy numbers have already been successfully correlated with the protein levels of humans and primates, it would perhaps be useful to ascertain copy number variation at the pancreatic amylase locus. Though the genome browser reference sequence shows two copies of the pancreatic amylase in the human genome, the assembly has undergone considerable change since March 2006 particularly for amylase gene cluster thus highlighting the uncertainty associated with this region.

CHAPTER 4: BIBLIOGRAPHY

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature, 290*, 457-465.

Armour, J. A. L., Sismani, C., Patsalis, P. C., & Cross, G. (2000). Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Research, 28*, 605-609.

Armour, J. A., Palla, R., Zeeuwen, P. L., den Heijer, M., Schalkwijk, J., & Hollox, E. J. (2007). Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Research, 35*, e19.

Aitman, T. J., Dong, R., Vyse, T. J., Norsworthy, P. J., Johnson, M. D., Smith, J., et al.(2006). Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature, 439*, 851-855.

Babushok, D. V., & Kazazian, H. H., Jr. (2007). Progress in understanding the biology of the human mutagen LINE-1. *Human Mutation, 28*, 527-539.

Bailey, J. A., & Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews Genetics, 7*, 552-564.

Bloom, G. D., Carlsoo, B., & Danielsson, A. (1975). Dopamine-Induced Amylase Secretion from Guinea-Pig Submandibular-Gland. *British Journal of Pharmacology, 54*(4),523-528.

Brennan, C., Zhang, Y. Y., Leo, C., Feng, B., Cauwels, C., Aguirre, A. J., et al. (2004). High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Research, 64*, 4744-4748.

Britten, R. J., & Kohne, D. E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science, 161*, 529-540.

Bruder, C. E. G., Piotrowski, A., Gijsbers, A. A. C. J., Andersson, R., Erickson, S., de Stahl, T. D., et al(2008). Phenotypically concordant and discordant monozygotic twins display

different DNA copy number variation profiles. *American Journal of Human Genetics, 82*, 763-771.

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., et al. (2003).Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America, 100*, 5280 5285.

Brown, N. M., S. Bernacki, et al. (2001). Fluorescent, multiplexed, automated, primer-extension assay for 3120+1G-->A and I148T mutations in cystic fibrosis. *Clinical Chemistry* **,** 2053-5.

Carvalho, B., Ouwerkerk, E., Meijer, G. A., & Ylstra, B. (2004). High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *Journal of Clinical Pathology, 57*, 644-646.

Caspersson, T., Farber, S., Foley, G. E., Kudynowski, J., Modest, E. J., Simonsson, E., et al. (1968).Chemical differentiation along metaphase chromosomes. *Experimental Cell Research, 49*, 219-222.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics, 37*, 1243-1246.

Clayton, D. A. (1992). Transcription and replication of animal mitochondrial DNAs. *International Review of Cytology*, *141*, 217 232.

Cost, G. J., Feng, Q., Jacquier, A., & Boeke, J. D. (2002). Human L1 element target-primed reverse transcription in vitro. *European Molecular Biology Organisation Journal, 21*, 5899-5910.

Craig, J. M., & Bickmore, W. A. (1993). Chromosome bands--flavours to savour. *Bioessays, 15*,349-354.

Darai-Ramqvist, E., Sandlund, A., Muller, S., Klein, G., Imreh, S., & Kost-Alimova, M. (2008).Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. *Genome Research, 18*, 370 379.

De Cid, R., E. Riveira-Munoz, et al. (2009). Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nature Genetics* , 41, 211-5.

Doi, S., Tomita, N., Higasiyama, M., Yokouchi, H., Horii, A., Yasuda, T., et al. (1991). Expression of Alpha Amylase Isozymes in Human Thyroid Tissues. *Cancer Research,51*, 3544-3549.

Dopman, E. B. and D. L. Hartl (2007). A portrait of copy-number polymorphism in Drosophila melanogaster.*Proceedings of the National Academy of Sciences of the United States of America* 104, 19920-5.

Dracopoli, N. C., & Meisler, M. H. (1990). Mapping the Human Amylase Gene-Cluster on the Proximal Short Arm of Chromosome-1 Using a Highly Informative (Ca)N Repeat. *Genomics, 7*, 97-102.

Emi, M., Horii, A., Tomita, N., Nishide, T., Ogawa, M., Mori, T., et al. (1988). Overlapping 2 Genes in Human DNA - a Salivary Amylase Gene Overlaps with a Gamma- Actin Pseudogene That Carries an Integrated Human Endogenous Retroviral DNA. *Gene, 62*, 229-235.

Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L., et al. (2007). FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature Genetics, 39*, 721-723.

Feng, Q., Moran, J. V., Kazazian, H. H., Jr., & Boeke, J. D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell, 87*, 905-916.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics, 7*, 85-97.

Fontaine, K. M., J. R. Cooley, et al. (2007). Evidence for paternal leakage in hybrid periodical cicadas (Hemiptera: Magicicada spp. *Public Library of Science One*, 2, e892.

Ford, C. E. & Hamerton, J. L. (1956) The chromosomes of man. *Nature* 178, 1020–1023.

Fridlyand, J., Snijders, A. M., Ylstra, B., Li, H., Olshen, A., Segraves, R., et al. (2006). Breast tumor copy number aberration phenotypes and genomic instability. *BioMed Central Cancer, 6*, 96.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science, 307*, 1434-1440.

Goodier, J. L., Ostertag, E. M., Du, K., & Kazazian, H. H., Jr. (2001). A novel active L1 retrotransposon subfamily in the mouse. *Genome Research, 11*, 1677-1685.

Groot, P. C., Bleeker, M. J., Pronk, J. C., Arwert, F., Mager, W. H., Planta, R. J., et al.(1989). The Human Amylase Multigene Family Consists of Haplotypes with Variable Numbers of Genes. *Cytogenetics and Cell Genetics, 51*, 1009-1009.

Groot, P. C., Mager, W. H.,& Frants, R. R. (1991). Interpretation of Polymorphic DNA Patterns in the Human Alpha-Amylase Multigene Family. *Genomics, 10*, 779-785.

Gumucio, D. L., Wiebauer, K., Caldwell, R. M., Samuelson, L. C., & Meisler, M. H. (1988). Concerted Evolution of Human Amylase Genes. *Molecular and Cellular Biology, 8*, 1197-1205.

Hayashi, Y., Fukayama, M., Koike, M., & Nakayama, T. (1986). Amylase in Human Lungs and the Female Genital-Tract - Histochemical and Immunohistochemical Localization. *Histochemistry, 85*, 491-496.

Heid, C. A., Stevens, J., Livak, K. J., & Williams, P. M. (1996). Real time quantitative PCR. *Genome Research, 6*, 986-994.

Higgs, D. R., Vickers, M. A., Wilkie, A. O., Pretorius, I. M., Jarman, A. P., & Weatherall, D. J. (1989). A review of the molecular genetics of the human alpha-globin gene cluster. *Blood, 73*, 1081-1104.

Higuchi, R., Dollinger, G., Walsh, P. S., & Griffith, R. (1992). Simultaneous amplification and detection of specific DNA sequences. *Biotechnology (N Y), 10*, 413-417.

Higuchi, R., Fockler, C., Dollinger, G., & Watson, R. (1993). Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biotechnology (N Y), 11*, 1026-1030.

Hjorth, J. P., Meisler, M., & Nielsen, J. T. (1979). Genetic-Variation in Amount of Salivary Amylase in the Bank Vole, Clethrionomys-Glareola. *Genetics, 92*, 915-930.

Hollox, E. J., Detering, J. C., & Dehnugara, T. (2009). An integrated approach for measuring copy number variation at the FCGR3 (CD16) locus. *Human Mutation, 30*(3), 477-484.

Jung, R., Soondrum, K., & Neumaier, M. (2000). Quantitative PCR. *Clinical Chemistry and Laboratory Medicine, 38*, 833-836.

Horii, A., Emi, M., Tomita, N., Nishide, T., Ogawa, M., Mori, T., et al. (1987). Primary Structure of Human Pancreatic Alpha-Amylase Gene - Its Comparison with Human Salivary Alpha-Amylase Gene. *Gene, 60*, 57-64.

Hsu, T. C. Human and Mammalian Cytogenetics: an Historical Perspective (Springer, New York, 1979).

Jones, J. M., Keller, S. A., Samuelson, L. C., Osborn, L., Rosenberg, M. P., & Meisler, M. H. (1989). A Salivary Amylase Transgene Is Efficiently Expressed in Liver but Not in Parotid-Gland of Transgenic Mice. *Nucleic Acids Research, 17*, 6613-6623.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., et al. (1992). Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors. *Science, 258*, 818-821.

Kallioniemi, O. P., Kallioniemi, A., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., et al. (1993). Comparative Genomic Hybridization - a Rapid New Method for Detecting and Mapping DNA Amplification in Tumors. *Seminars in Cancer Biology, 4*, 41-46.

Kondo, R., E. T. Matsuura, et al. (1992). Further observation of paternal transmission of Drosophila mitochondrial DNA by PCR selective amplification method. *Genetics Research* 59, 81-4.

Koszul, R., & Fischer, G. (2009). A prominent role for segmental duplications in modeling Eukaryotic genomes. *Comptes Rendus Biologies, 332*, 254-266.

Lam, K. W., & Jeffreys, A. J. (2006). Processes of copy-number change in human DNA: the dynamics of {alpha}-globin gene deletion. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 8921-8927.

Lam, K. W. G., & Jeffreys, A. J. (2007). Processes of de novo duplication of human alpha-globin genes. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 10950-10955.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*, 860-921.

Landegent, J. E. *et al.* (1985) Chromosomal localization of a unique gene by non autoradiographic *in situ* hybridization. *Nature* 317, 175–177.

Lavie, L., Maldener, E., Brouha, B., Meese, E. U., & Mayer, J. (2004). The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Research, 14*, 2253-2260.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., et al. (2007). The diploid genome sequence of an individual human. *Public library of science Biology, 5*, 2113-2144.

Lee, J. A., Carvalho, C. M., & Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell, 131*, 1235-1247.

Lieber, M. R., Lu, H., Gu, J., & Schwarz, K. (2008). Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate nonhomologous DNA end joining: relevance to cancer, aging, and the immune system. *Cell Research, 18*, 125-133.

Lieber, M. R., Ma, Y. M., Pannicke, U., & Schwarz, K. (2003). Mechanism and regulation of human non homologous DNA end-joining. *Nature Reviews Molecular Cell Biology, 4*, 712-720.

Locke, D. P., A. J. Sharp, et al. (2006). Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *American Journal of Human Genetics* 79, 275-90.

Lower, K. M., Hughes, J. R., De Gobbi, M., Henderson, S., Viprakasit, V., Fisher, C., et al. (2009). Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 21771-21776.

Lupski, J. R., & Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *Public library of science Genetics, 1*, e49.

Mau, M., Sudekum, K. H., Johann, A., Sliwa, A., & Kaiser, T. M. (2010). Indication of higher salivary alpha amylase expression in hamadryas baboons and geladas compared to chimpanzees and humans. *Journal of Medical Primatology, 39*, 187-190.

Mamtani, M., Rovin, B., Brey, R., Camargo, J. F., Kulkarni, H., Herrera, M., et al. (2008). CCL3L1 gene-containing segmental duplications and polymorphisms in CCR5 affect risk of systemic lupus erythaematosus. *Annals of the Rheumatic Diseases, 67*, 1076-1083.

McCarroll, S. A. (2008). Copy-number analysis goes more than skin deep. *Nature Genetics, 40*, 56.

McCarroll, S. A., & Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature Genetics, 39*(7 Suppl), S37-42.

McKinney, C., Merriman, M. E., Chapman, P. T., Gow, P. J., Harrison, A. A., Highton, J., et al. (2008). Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Annals of the Rheumatic Diseases, 67*(3), 409-413.

Merritt, A. D., & Karn, R. C. (1977). The human alpha-amylases. *Advances in Human Genetics, 8*,135-234.

Merritt, A. D., Rivas, M. L., Bixler, D., & Newell, R. (1973). Salivary and Pancreatic Amylase Electrophoretic Characterizations and Genetic Studies. *American Journal of Human Genetics, 25*, 510-522.

Meusel, M. S. and R. F. Moritz (1993). Transfer of paternal mitochondrial DNA during fertilization of honeybee (Apis mellifera L.) eggs. *Current Genetics*, 24, 539-43.

Munke, M., Lindgren, V., de Martinville, B., & Francke, U. (1984). Comparative analysis of mouse human hybrids with rearranged chromosomes 1 by in situ hybridization and Southern blotting: high resolution mapping of NRAS, NGFB, and AMY on human chromosome 1. *Somatic Cell and Molecular Genetics, 10*, 589-599.

Nishide, T., Nakamura, Y., Emi, M., Yamamoto, T., Ogawa, M., Mori, T., et al. (1986). Primary Structure of Human Salivary Alpha-Amylase Gene. *Gene, 41*, 299-304.

Pardue, M. L. & Gall, J. G. (1969). Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America* 64, 600–604.

Patsalis, P. C., Kousoulidou, L., Sismani, C., Mannik, K., & Kurg, A. (2005). MAPH: from gels to microarrays. *European Journal of Medical Genetics, 48*, 241-249.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics, 39*, 1256-1260.

Pinkel, D., Seagraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics, 20*, 207-211.

Piotrowski, A., Bruder, C. E. G., Andersson, R., de Stahl, T. D., Menzel, U., Sandgren, J., et al. (2008). Somatic mosaicism for copy number variation in differentiated human tissues. *Human Mutation, 29*, 1118 1124.

Ravetch, J. V., & Perussia, B. (1989). Alternative membrane forms of Fc gamma RIII(CD16) on human natural killer cells and neutrophils. Cell type-specific expression of two genes that differ in single nucleotide substitutions. The *Journal of Experimental Medicine, 170*, 481-497.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature, 444*, 444-454.

Samonte, R. V., & Eichler, E. E. (2002). Segmental duplications and the evolution of the primate genome. *Nature Reviews Genetics, 3*, 65-72.

Samuelson, L. C., Wiebauer, K., Gumucio, D. L., & Meisler, M. H. (1988). Expression of theHuman Amylase Genes - Recent Origin of a Salivary Amylase Promoter from an Actin Pseudogene. *Nucleic Acids Research, 16*, 8261-8276.

Samuelson, L. C., Wiebauer, K., Snow, C. M., & Meisler, M. H. (1990). Retroviral and Pseudogene Insertion Sites Reveal the Lineage of Human Salivary and Pancreatic Amylase Genes from a Single Gene during Primate Evolution. *Molecular and Cellular Biology, 10*, 2513-2520.

Schibler, U., Pittet, A. C., Young, R. A., Hagenbuchle, O., Tosi, M., Gellman, S., et al. (1982). The Mouse Alpha-Amylase Multigene Family - Sequence Organization of Members

Expressed in the Pancreas, Salivary Gland and Liver. *Journal of Molecular Biology, 155*, 247-266.

Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F., & Pals, G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research, 30*

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., et al. (2004). Large-scale copy number polymorphism in the human genome. *Science, 305*, 525-528.

Sellner, L. N., & Taylor, G. R. (2004). MLPA and MAPH: New techniques for detection of gene deletions. *Human Mutation, 23*, 413-419.

Shear, M., Gibson, S., & Vanderme.E. (1973). Qualitative and Quantitative Histochemical Investigation of Effects of Pilocarpine Stimulation on Amylase Levels in Rat Submaxillary and Parotid-Glands. *Journal of Dental Research, 52*, 619-619.

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends in Genetics, 18*, 74-82.

Swergold, G. D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. *Molecular and Cellular Biology, 10*, 6718-6729.

The International HapMap Consortium. The International HapMap Project. Nature 426, 789-796 (2003).

Tjio, H. J. & Levan, A. (1956)The chromosome numbers of man. *Hereditas* 42, 1–6.

Trask, B. J. (1991)Fluorescence *in situ* hybridization: applications in cytogenetics and gene mapping. *Trends in Genetics.* 7, 149–154.

Ting, C. N., Rosenberg, M. P., Snow, C. M., Samuelson, L. C., & Meisler, M. H. (1992).Endogenous Retroviral Sequences Are Required for Tissue-Specific Expression of a Human Salivary Amylase Gene. *Genes & Development, 6*, 1457-1465.

Tomita, N., Matsuura, N., Horii, A., Emi, M., Nishide, T., Ogawa, M., et al. (1988). Expression of Alpha-Amylase in Human-Lung Cancers. *Cancer Research, 48*, 3292-3296.

Townson, J. R., Barcellos, L. F., & Nibbs, R. J. (2002). Gene copy number regulates the production of the human chemokine CCL3-L1. *European Journal of Immunology, 32*, 3016-3026.

Tricoli, J. V., & Shows, T. B. (1984). Regional Assignment of Human Amylase (Amy) to P22-]P21 of Chromosome-1. *Somatic Cell and Molecular Genetics, 10*, 205-210.

Van den Engh, G., Sachs, R. & Trask, B. J. (1992)Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science* 257, 1410–1412.

Van Ommen, G. J. B. (2005). Frequency of new copy number variation in humans. *Nature Genetics, 37*, 333-334.

Vinckenbosch, N., I. Dupanloup, et al. (2006). Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of National Acadademy of Science of the United States of America* 103, 3220-5.

Vogel, F. & Motulsky, A. G. *Human Genetics: Problems and Approaches* (Springer, Berlin, 1997)

Walker, S., Janyakhantikul, S., & Armour, J. A. (2009). Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs. *Genomics, 93*, 98-103.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., et al.(2002). Initial sequencing and comparative analysis of the mouse genome. *Nature, 420*, 520-562.

Wei, W., Gilbert, N., Ooi, S. L., Lawler, J. F., Ostertag, E. M., Kazazian, H. H., et al. (2001). Human L1 retrotransposition: cis preference versus trans complementation. *Molecular and Cellular Bioliology, 21*, 1429-1439.

Wiegant, J. *et al.* (1992) High-resolution *in situ* hybridization using DNA halo preparations. *Human Molecular Genetics* 1, 587–591.

Whitten, R. O., Chandler, W. L., Thomas, M. G. E., Clayson, K. J., & Fine, J. S. (1988). Survey of Alpha-Amylase Activity and Isoamylases in Autopsy Tissue. *Clinical Chemistry, 34*, 1552-1555.

Wiebauer, K., Gumucio, D. L., Jones, J. M., Caldwell, R. M., Hartle, H. T., & Meisler, M. H. (1985). A 78-Kilobase Region of Mouse Chromosome-3 Contains Salivary and Pancreatic Amylase Genes and a Pseudogene. *Proceedings of the National Academy of Sciences of the United States of America, 82*, 5446-5449.

Winderickx, J., Battisti, L., Motulsky, A. G., & Deeb, S. S. (1992). Selective expression of human X chromosome-linked green opsin genes. *Proceedings of the National Academy of Sciences of the United States of America, 89*(20), 9710-9714.

Zabel, B. U., Naylor, S. L., Sakaguchi, A. Y., Bell, G. I., & Shows, T. B. (1983). High-Resolution Chromosomal Localization of Human Genes for Amylase, Proopiomelanocortin, Somatostatin, and a DNA Fragment (D3s1) by Insitu Hybridization. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences, 80*(22), 6932-6936.

Zakowski, J. J., Gregory, M. R., & Bruns, D. E. (1984). Amylase from Human Serous Ovarian-Tumors - Purification and Characterization. *Clinical Chemistry, 30*, 62-68.

Zhang, F., Gu, W. L., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics, 10*, 451-481.

Zhao, X. J., Li, C., Paez, J. G., Chin, K., Janne, P. A., Chen, T. H., et al. (2004). An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research, 64*, 3060-3071.

Zhou, X. F., Mok, S. C., Chen, Z., Li, Y., & Wong, D. T. W. (2004). Concurrent analysis of loss of heterozygosity (LOH) and copy number abnormality (CNA) for oral premalignancy progression using the Affymetrix 10K SNP mapping array. *Human Genetics, 115*, 327-330.